Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Neurosymbolic Causally-Grounded Perception-Action Architecture for Artificial Embodied Agents

Master's Thesis

2026
Bc. Miroslav Cibula

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

Neurosymbolic Causally-Grounded Perception-Action Architecture for Artificial Embodied Agents

Master's Thesis

Study Programme:	Cognitive Science
Field of Study:	Computer Science
Department:	Department of Applied Informatics
Supervisor:	prof. Ing. Igor Farkaš, Dr.

Bratislava, 2026
Bc. Miroslav Cibula

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Miroslav Cibula
Študijný program: kognitívna veda (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Neurosymbolic Causally-Grounded Perception-Action Architecture for Artificial Embodied Agents
Neurosymbolová kauzálne ukotvená senzomotorická architektúra pre umelé vtelené agenty

Anotácia: Znalosť kauzálnych vzťahov je nevyhnutná pre robustné a flexibilné učenie a uvažovanie u ľudí. Kognitívne reprezentácie sprostredkúvajúce tieto operácie sú ukotvené v percepcii a akcii, kontextualizované, hierarchicky usporiadané vo viacerých úrovniach abstrakcie a vyvíjajú sa postupne počas života agenta. Aplikácia týchto mechanizmov vo vtelených agentoch poskytuje základ pre komplexnejšie kognitívne operácie mimo čisto fyzickej domény.

Cieľ: 1. Navrhnite a implementujte neurosymbolový, bio-plauzibilný perceptuálny systém spracúvajúci vizuálnu informáciu získanú z dynamických scén a konštruujúci hybridné konceptuálne reprezentácie.
2. Navrhnite a implementujte komplementárny usudzovací a akčný systém s aktívnou kauzálnou inferenciou na učenie sa a operovanie nad hierarchickými abstraktnými reprezentáciami generovanými perceptuálnym systémom.
3. Zanalyzujte kritické subsystémy implementovanej časti architektúry.

Literatúra: Butz, M. V. et al. (2025). Contextualizing predictive minds. Neuroscience & Biobehavioral Reviews, 168, 105948.
Gopnik, A., & Schulz, L. E. (Eds.). (2007). Causal Learning: Psychology, Philosophy, and Computation. Oxford University Press.
Toth, C. et al. (2022). Active Bayesian causal inference. In Advances in Neural Information Processing Systems (Vol. 35, pp. 16261–16275).

Vedúci: prof. Ing. Igor Farkaš, Dr.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: doc. RNDr. Tatiana Jajcayová, PhD.

Dátum zadania: 28.02.2026

Dátum schválenia: 17.03.2026

prof. Ing. Igor Farkaš, Dr.
garant študijného programu

.....
študent

.....
vedúci práce

Comenius University Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname:	Bc. Miroslav Cibula
Study programme:	Cognitive Science (Single degree study, master II. deg., full time form)
Field of Study:	Computer Science
Type of Thesis:	Diploma Thesis
Language of Thesis:	English
Secondary language:	Slovak

Title: Neurosymbolic Causally-Grounded Perception-Action Architecture for Artificial Embodied Agents

Annotation: Understanding causal relationships in the world is principal to robust and flexible human learning and reasoning. Cognitive representations facilitating these operations are multi-modally grounded in perception and action, heavily contextualized, hierarchically organized, incorporating multiple levels of abstraction, and developed progressively over the life of an agent. When deployed in an embodied agent, such mechanisms will provide the foundation for more complex cognitive operations beyond the purely physical domain.

Aim: 1. Design and develop a transparent, neurosymbolic, biologically plausible perceptual system able to process visual information from dynamic scenes to construct hybrid and conceptual representations.
2. Design and implement a complementary reasoning and action-generating system leveraging active causal inference to learn and operate on hierarchically-structured abstract representations yielded by the perceptual system.
3. Analyze critical subsystems of the implemented part of the architecture.

Literature:
Butz, M. V. et al. (2025). Contextualizing predictive minds. Neuroscience & Biobehavioral Reviews, 168, 105948.
Gopnik, A., & Schulz, L. E. (Eds.). (2007). Causal Learning: Psychology, Philosophy, and Computation. Oxford University Press.
Toth, C. et al. (2022). Active Bayesian causal inference. In Advances in Neural Information Processing Systems (Vol. 35, pp. 16261–16275).

Supervisor:	prof. Ing. Igor Farkaš, Dr.
Department:	FMFI.KAI - Department of Applied Informatics
Head of department:	doc. RNDr. Tatiana Jajcayová, PhD.
Assigned:	28.02.2026
Approved:	17.03.2026

prof. Ing. Igor Farkaš, Dr.
Guarantor of Study Programme

.....
Student

.....
Supervisor

Acknowledgments

First of all, I would like to sincerely thank my primary supervisor, Prof. Dr. Igor Farkaš, as well as my (unofficial) secondary advisor, Dr. Kristína Malinovská, for their constant guidance, support, and mentorship throughout my modest academic journey (and beyond it), including this thesis. The present project accumulates the results of multiple research works spanning several years, and therefore, I would like to express my gratitude to the researchers I was fortunate to interact and collaborate with, who significantly contributed with their advice, ideas, and technical or research work to this thesis. In no specific order:

- Dr. Matthias Kerzel (*Department of Informatics, University of Hamburg*), for our multiple collaborations in sensorimotor learning research and for his useful feedback, which has been partially used here;
- Prof. Dr. Markus Vincze (*Automation and Control Institute, TU Wien*), for his advice on the computer vision aspects used here, and for guidance during the development of the perceptual subsystem;
- Dr. Michal Vavrečka (*Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague*), for his help with my previous research underlying this thesis, and our discussions on various aspects of robotic sensorimotor learning;
- Dr. Loïc Goasguen and Prof. Dr. Matěj Hoffmann (*Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague*), for their advice and feedback during the earlier development of the world model presented in this thesis;
- Prof. Dr. Robert Peharz (*Institute of Machine Learning and Neural Computation, TU Graz*), for his useful pointers and advice toward several aspects of causal and Bayesian learning used here;
- Prof. Dr. Mohan Sridharan (*School of Informatics, University of Edinburgh*), for our helpful discussions on neurosymbolic and hierarchical systems in the earlier stages of this project, and for his pointers toward several representational aspects of this work.

I would like to thank my two very good friends, Aly Debevec-Kruse and Ján Hučko, for reviewing parts of this text to make them (hopefully) human-readable, for their useful advice and feedback on various aspects of this thesis, and for their support. Finally, I would also like to thank the members of the *Cognition and Neural Computation* research group for their feedback and helpful advice on multiple occasions.

Various parts of this project, including the preliminary research, were supported by the Horizon Europe project TERAIS, GA no. 101079338, and the Slovak research projects VEGA 1/0373/23, KEGA 022UK-4/2023, and APVV 21-0105. The presented research results were in part obtained using the high-performance computational resources of HPC Devana provided by the national EU-funded project no. 311070AKF2 National Competence Centre for HPC, and of HPC Perun provided by the Supercomputing Center at TU Košice, with support from the EU, funds of the Recovery and Resilience Plan of the Slovak Republic, project no. 17I03-04 P03-00001.

Abstract

Achieving grounded, common-sense cognition in artificial agents requires integrating perception, action, and learning into an embodied interaction loop. Developmental accounts suggest that before language and explicit symbolic reasoning, infants acquire early causal and object-related knowledge through sensorimotor exploration. Modeling this stage computationally requires bridging non-symbolic perceptual processing with structured reasoning: purely neural models often lack transparency and explicit relational structure, while classical symbolic approaches struggle with noisy, high-dimensional perceptual data. This thesis, therefore, proposes a minimal neurosymbolic perception-action architecture enabling an embodied agent to construct grounded perceptual representations and learn structured sensorimotor world models. The architecture integrates object-centric visual perception, conceptual representation, and causal world modeling. Visual observations are transformed into structured object records, retained object-local evidence, visual and kinematic descriptors, and conceptual-space projections. Sensorimotor regularities are represented in a structured world model inspired by structural causal modeling, where variables and local mechanisms make action-effect relations explicit and inspectable. The world model is then queried for prediction, target-conditioned action inference, and curiosity-conditioned target selection, providing a reduced implementation of inference-driven action generation. The thesis primarily focuses on the theoretical design of the architecture and also provides a proof-of-concept implementation of selected components. The perceptual experiments evaluate whether real RGB-D object observations can be converted into object-centric conceptual entities and structured scene descriptors, while the world-model experiments evaluate whether a minimal differentiable causal graph trained on proprioceptive motor-babbling data can support forward prediction, model-consistent action inference, and internally generated target pursuit. The two subsystems are formulated within a common perception-action setting, but they are not yet evaluated as a single closed-loop multimodal agent. Overall, the results support the feasibility of the proposed representational decomposition under restricted conditions and identify the empirical integration of visual object-level variables with learned sensorimotor mechanisms as the central direction for future work.

Keywords: perception-action loop, causal learning, neurosymbolic systems, conceptual spaces, embodied cognition

Abstrakt

Dosiahnutie ukotvenej, vysokoúrovňovej kognície u umelých agentov vyžaduje integráciu percepcie, akcie a učenia do perceptuálno-motorickej slučky. Vývinové prístupy naznačujú, že pred nadobudnutím jazyka a explicitného symbolického usudzovania si deti osvojujú rané kauzálne poznatky a poznatky o objektoch prostredníctvom senzomotorického skúmania. Výpočtové modelovanie tohto štádia vyžaduje premostenie nesymbolového perceptuálneho spracovania so štruktúrovaným usudzovaním: čisto neurónovým modelom často chýba transparentnosť a explicitná relačná štruktúra, zatiaľ čo klasické symbolové prístupy majú problém so nepresnými, vysokodimenzionálnymi surovými perceptálnymi dátami. Táto práca preto navrhuje minimálnu neurosymbolovú perceptuálno-motorickú architektúru, ktorá stelesnenému agentovi umožňuje konštruovať ukotvené perceptuálné reprezentácie a učiť sa štruktúrované senzomotorické modely sveta. Architektúra integruje objektovo-orientovanú vizuálnu percepciu, konceptuálne reprezentácie a kauzálne modelovanie sveta. Vizuálne percepty sú transformované na štruktúrované reprezentácie objektov, vzhľadové a kinematické deskriptory a projekcie do konceptuálnych priestorov. Senzomotorické regularity sú reprezentované v štruktúrovanom modeli sveta inšpirovanom štrukturálnym kauzálnym modelovaním, v ktorom premenné a lokálne mechanizmy explicitne ukotvujú vzťahy typu akcia–efekt. Takýto model je následne použitý na predikciu, cieľom podmienenú inferenciu akcií a exploratívne správanie založené na vnútornej motivácii. Práca sa primárne zameriava na teoretický návrh architektúry a zároveň poskytuje proof-of-concept implementáciu vybraných komponentov. Perceptuálne experimenty testujú, či je možné reálne RGB-D pozorovania dynamických scén previesť na objektovo-orientované konceptuálne entity a štruktúrované deskriptory scény, zatiaľ čo experimenty so senzomotorickým modelom sveta testujú, či minimálny diferencovateľný kauzálny model trénovaný na proprioceptívnych dátach zo simulovaného motorického bľabotania dokáže podporovať doprednú predikciu, modelovo konzistentnú inferenciu akcií a interne generované sledovanie cieľov. Oba subsystémy sú formulované v spoločnom perceptuálno-motorickom rámci, zatiaľ však nie sú implementované ako plne uzavretá multimodálna slučka pre vtelených agentov. Celkovo výsledky podporujú uskutočniteľnosť navrhovanej reprezentačnej dekompozície v obmedzených podmienkach a identifikujú empirickú integráciu vizuálnych premenných na úrovni objektov s naučenými senzomotorickými mechanizmami ako hlavný smer budúceho výskumu.

Kľúčové slová: perceptuálno-motorická slučka, kauzálne učenie, neurosymbolové systémy, konceptuálne priestory, vtelená kognícia

Contents

Introduction **1**

1 Preliminaries **5**
1.1 Sensorimotor Cognition 5
1.1.1 Cognitive Perspective 6
1.1.2 Computational Perspective 9
1.2 Visual Perception 11
1.2.1 Neurocognitive Correlates 11
1.2.2 Active Computer Vision 14
1.2.3 Conceptual Spaces 15
1.3 Learning and Reasoning 18
1.3.1 Bayesian Learning and Probabilistic Graphical Models 18
1.3.2 Causal Inference and Autonomous Experimental Design 20
1.3.3 Sequential Decision-Making and Active Inference 22
1.3.4 Neurosymbolic and Multilevel Architectures 24

2 Related Work **27**
2.1 Sensorimotor World Models 27
2.2 Computational Visual Perception 29

3 Aims and Problem Formulation **31**
3.1 Scope 31
3.2 Architectural Desiderata 32
3.3 Problem Formalization 34

4 Proposed Architecture **37**
4.1 Overall System Architecture 37
4.2 Visual Perception and Semantic Representations 39
4.2.1 Formalization 41
4.2.2 Subsystems 42
4.2.2.1 Input Discretization 42
4.2.2.2 Perceptual Memory and Maintenance 44
4.2.2.3 Physical Behavior Monitoring 45

4.2.2.4 Visual Quantization 47
4.2.2.5 Scene Composition Parsing 48
4.2.2.6 Conceptual Representation Construction 50
4.2.3 Neurocognitive Correlates 51
4.3 Neurosymbolic Knowledge Representation 54
4.3.1 Model Structure 54
4.3.2 Causal Learning and Discovery 57
4.3.3 Intrinsic Motivation Generation 58
4.3.4 State Inference 59
4.3.5 Action Generation 62
4.3.6 Meta-Learning 64

5 Experiments and Results **65**
5.1 Perceptual Results 65
5.1.1 Implementation Overview 65
5.1.2 Experiment A1: Kinematic Concept Formation 67
5.1.3 Experiment A2: Visual Concept Formation 70
5.1.4 Experiment A3: Multidomain Representation 71
5.1.5 Structured Scene Descriptor 72
5.2 World Model Results 73
5.2.1 Experimental Configuration 74
5.2.2 Experiment B1: Sensorimotor Transition Prediction 76
5.2.3 Experiment B2: Action Inference 77
5.2.4 Experiment B3: Curiosity-Conditioned Target Selection 77

6 Discussion **81**
6.1 Perception 82
6.2 World Model 84
6.3 Perception-Action Integration 86
6.4 Extensions 87
6.4.1 World Model Hierarchization 87
6.4.2 World Model Contextualization 89
6.4.3 Artificial Formal Language of Thought 90
6.4.4 Natural Language Emergence 92
6.4.5 Sensorimotor Social Cognition 93

Conclusion **95**

A Implementation Source Code **121**

List of Figures

1.1 High-level schema of a generic perception-action loop 8
1.2 High-level diagram of the early visual processing pathway 12
1.3 Neural anatomy of the ventral and dorsal visual pathways 13
1.4 Elementary topologies in Bayesian networks 19

4.1 High-level organization of the proposed architecture 38
4.2 Full high-level diagram of the proposed perceptual architecture 40
4.3 Functional neurocognitive interpretation of the perceptual system . . . 53
4.4 World-model variable types according to their topological role 56

5.1 Real-world experimental setup for the perceptual experiments 68
5.2 3D trajectories of inspected motion records 69
5.3 Motion conceptual-space projection of exported trajectories 71
5.4 Visual conceptual-space projection 73
5.5 Graph of a simple proprioceptive world model 75
5.6 Performance of the action inference for proprioceptive targets 78
5.7 Curiosity targets and predicted proprioceptive outcomes 79

List of Tables

5.1 Objects and variants used in the perceptual experiments 66
5.2 Configuration of evaluated perceptual representations 67
5.3 Motion trajectory quality summary 70
5.4 Visual records and prototype distances 72
5.5 Multi-space conceptual entities 72
5.6 Schema of the real-time scene descriptor 74
5.7 Predictive performance on test motor-babbling transitions 76

List of Algorithms

4.1 World-model forward query as mental simulation 60
4.2 Gradient-based action generation through the world model 63

List of Abbreviations

ABCI	active Bayesian causal inference
ACV	active computer vision
AIT	anterior inferotemporal cortex
ATL	anterior temporal lobe
BED	Bayesian experimental design
BN	Bayesian network
CRs	circular reactions
DAG	directed acyclic graph
DCT	discrete cosine transform
DoF	degrees of freedom
DVP	dorsal visual pathway
EEMS	ephemeral eidetic memory storage
EFE	expected free energy
FM	forward model
GD	gradient descent
HOEL	higher object-event layer
IBEL	intermediate body-event layer
IM	inverse model
IoM	intersection over minimum
IoU	intersection over union
ITC	inferotemporal cortex
LGN	lateral geniculate nucleus
LoT	language of thought
LSML	lower sensorimotor layer
MAE	mean absolute error
MDP	Markov decision process
MLP	multilayer perceptron
MSE	mean squared error
MST	medial superior temporal area
MT	middle temporal visual area
PAL	perception-action loop
PCA	principal component analysis
PGM	probabilistic graphical model
PIT	posterior inferotemporal cortex
PnP	perspective-n-point
POMDP	partially-observable Markov decision process
PPC	posterior parietal cortex
QSR	qualitative spatial reasoning
RGC	retinal ganglion cell
RL	reinforcement learning
RV	random variable
SCM	structural causal model
SMCs	sensorimotor contingencies
V1	primary visual cortex
VFE	variational free energy
VVP	ventral visual pathway
WM	world model

Summary of Notation

Capital letters (e.g., X) are used for random variables, sets, or matrices if boldfaced (e.g., $\boldsymbol{M}$). Lower-case letters denote scalars (e.g., x), and the values of random variables (e.g., $X = x$). Vectors are denoted by lower-case bold letters (e.g., $\boldsymbol{v}$).

General Notation

$\mathbb{R}$	real numbers
$\mathbb{N}$	non-negative integers
$\mathbb{N}^+$	positive integers
$1\!:\!n$	index interval $1, \ldots, n$
$\boldsymbol{u} \subseteq \boldsymbol{v}$	subvector of $\boldsymbol{v}$
$\boldsymbol{u} \subset \boldsymbol{v}$	proper subvector of $\boldsymbol{v}$
$\boldsymbol{u} \cup \boldsymbol{v}$	vector concatenation
$\|\boldsymbol{u}\|_p$	L^p-norm of $\boldsymbol{u}$
$\mathcal{A} \subseteq \mathcal{B}$	subspace/subset of $\mathcal{B}$
$\mathcal{A} \subset \mathcal{B}$	proper subspace/subset of $\mathcal{B}$
$\mathrm{SO}(n)$	special orthogonal group in n dimensions
$\nabla_{\boldsymbol{x}}$	gradient operator w.r.t. $\boldsymbol{x}$
$f(x)\|_{x=x_0}$	expression $f(x)$ evaluated at assignment $x = x_0$

Probabilistic Modeling

$\mathcal{U}[a, b]$	uniform distribution bounded by $[a, b]$
$\mathcal{N}(\mu, \sigma^2)$	normal distribution centered around μ with variance σ^2
$p(X)$	probability of X
$p(X \mid Y)$	conditional probability of X given Y
$X \perp\!\!\!\perp Y \mid Z$	X conditionally independent of Y given Z
$X \not\perp\!\!\!\perp Y \mid Z$	X not conditionally independent of Y given Z
$X \sim p$	X distributed w.r.t. distribution p
$\mathbb{E}[X]$	expected value of X
$\mathbb{E}_p[X]$	expected value of X w.r.t. distribution p
$H(X)$	entropy of X
$D_{\mathrm{KL}}(p \parallel q)$	Kullback-Leibler divergence from distribution p to q
$\mathcal{M}$	probabilistic model
$\mathcal{X}$	random variable (RV) set
$\mathcal{V}$	set of RVs endogenous to model
$\mathcal{U}$	set of RVs exogenous to model
$\mathcal{F}$	set of structural

	mechanisms
$\mathrm{do}(X = x)$	Pearlian interventional operator
ξ	Bayesian experiment
ξ^*	optimal Bayesian experiment
Ξ	experimental space
$F[q, \boldsymbol{o}]$	variational free energy of distribution q w.r.t. $\boldsymbol{o}$
$G(\pi)$	expected free energy of policy π
C	preference parameter

Graph Theory

$\mathcal{G}$	graph
$\mathcal{G}' \subseteq \mathcal{G}$	subgraph of $\mathcal{G}$
$\mathrm{pa}(x)$	parent nodes of x
$\deg^-(x)$	indegree of x

Environment Formalization

t	discrete environmental time step
Δt	discrete environmental time difference
h	finite time horizon
γ	discount factor
d_s	state dimensionality
d_q	action/proprioceptive dimensionality
c	visual history length
$\boldsymbol{s}(t)$	state vector at time t
$\boldsymbol{a}(t)$	action vector at time t
$\boldsymbol{o}(t)$	observation vector at time t
$\boldsymbol{v}(t)$	visual percept at time t
$\boldsymbol{q}(t)$	joint configuration at time t
$\boldsymbol{s}^{\mathrm{vis}}(t)$	visually inferable state subvector at time t
$\hat{\boldsymbol{s}}(t)$	estimated state vector at time t
$\hat{\boldsymbol{a}}(t)$	estimated action vector at time t
$\hat{\boldsymbol{o}}(t)$	estimated observation vector at time t
$\hat{\boldsymbol{s}}^{\mathrm{vis}}(t)$	estimated visual state descriptor at time t
π	policy
π^*	optimal policy
$\mathcal{E}$	environment
$\mathcal{S}$	state space
$\mathcal{A}$	action space
$\mathcal{O}$	observation space
$\mathcal{O}^{\mathrm{vis}}$	visual observation space
$\mathcal{Q}$	proprioceptive space
$\mathfrak{P}$	policy space
$\mathcal{H}_t$	structured sensorimotor history at time t
T_0	initial state distribution
$T(\boldsymbol{s'} \mid \boldsymbol{s}, \boldsymbol{a})$	state transition distribution
$R(\boldsymbol{s}, \boldsymbol{a})$	reward function
$\Omega(\boldsymbol{o} \mid \boldsymbol{s}, \boldsymbol{a})$	observation distribution
$\mathrm{FM}(\boldsymbol{s}, \boldsymbol{a})$	forward model
$\mathrm{IM}(\boldsymbol{s}, \boldsymbol{s'})$	inverse model

Conceptual Spaces

d	quality dimension
D	dimension set
δ	conceptual domain

$\mathcal{D}$	domain set
$\mathcal{C}$	conceptual space
$\mathcal{C}_J \subseteq \mathcal{C}$	conceptual subspace of $\mathcal{C}$ induced by domain subset J
$\boldsymbol{x}_\delta$	subvector of exemplar $\boldsymbol{x}$ restricted to dimensions of δ
w_δ	δ-specific salience weight
$\mathrm{dist}_\delta(\cdot,\cdot)$	δ-local within-domain metric
$\mathrm{dist}_{\mathcal{C}}(\cdot,\cdot)$	global metric of $\mathcal{C}$
ℓ	conceptual category
$\boldsymbol{\mu}_\ell$	centroid of ℓ
V_ℓ	region of ℓ

Architecture Components

$\boldsymbol{\theta}^{\mathrm{perc}}$	parameters of perceptual subsystem
Π	perceptual process
$\mathcal{M}^{\mathrm{wm}}(t)$	world model (WM) at time t
$\boldsymbol{\theta}^{\mathrm{wm}}$	parameters of WM subsystem

Perceptual Architecture

n_t	number of represented objects at time t
$\hat{\boldsymbol{s}}^{\mathrm{scene}}(t)$	scene-level visual descriptor at time t
$\hat{\boldsymbol{s}}_i^{\mathrm{obj}}(t)$	object descriptor of object i at time t
$\hat{\boldsymbol{s}}_i^{\mathrm{app}}(t)$	appearance descriptor of object i at time t
$\hat{\boldsymbol{s}}_i^{\mathrm{spat}}(t)$	spatial descriptor of object i at time t
$\hat{\boldsymbol{s}}_i^{\mathrm{kin}}(t)$	kinematic descriptor of object i at time t
$\boldsymbol{c}_i(t)$	conceptual representation of object i at time t
Σ	segmentation-tracking process
$\tilde{\mathcal{Z}}(t)$	candidate object-observation set at time t
$\mathcal{Z}(t)$	stabilized object-observation set at time t
$\iota_i(t)$	object identifier of object i at time t
$\boldsymbol{b}_i(t)$	bounding box of object i at time t
$\boldsymbol{m}_i(t)$	instance mask of object i at time t
$\omega_i(t)$	perceptual memory record of object i
$\Lambda_i(t)$	snapshot gallery of object i at time t
$\lambda_{i,k}$	k-th snapshot of object i
$\boldsymbol{\phi}_i(t)$	maintenance metadata of object i
$K_i(t)$	number of retained snapshots of object i
K_{eems}	maximum number of retained snapshots per object
BM	behavior-monitoring process
$\hat{\boldsymbol{p}}_i(t)$	estimated 6-DoF pose of object i
$\mathcal{T}_i(t)$	pose trajectory of object i up to time t
VQ	visual-quantization process
SCP	scene-composition

	process
$\hat{\boldsymbol{s}}_{ij}^{\mathrm{rel}}(t)$	pairwise relational descriptor between objects i and j
$J_i(t)$	domain subset available for object i at time t
$\mathcal{C}_i(t)$	object-specific conceptual subspace at time t
$\hat{\boldsymbol{s}}_{i,\delta}(t)$	δ-specific perceptual descriptor of object i
ζ_δ	δ-specific conceptualization map
$\boldsymbol{c}_{i,\delta}(t)$	conceptual point of object i in domain δ
$\boldsymbol{\mu}_{\ell,\delta}$	δ-specific prototype of category ℓ
$N_{\ell,\delta}$	number of exemplars assigned to category ℓ in domain δ

World Model

$p_t\left(\mathcal{U}^{\mathrm{wm}}\right)$	residual distribution of WM at time t
$\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t)$	graph induced by WM at time t
$\mathcal{V}^{\mathrm{act}}$	action/control-var. set
$\mathcal{V}^{\mathrm{obs}}$	observed-var. set
$\mathcal{V}^{\mathrm{lat}}(t)$	latent-var. set represented by WM at time t
$\mathcal{V}^{\mathrm{tar}}(t)$	target-var. set at time t
f_i	local structural mechanism for var./node X_i
U_i	exogenous residual var. for X_i
$u_i(t)$	residual value for X_i at time t
$\hat{p}_{X,t}(x)$	empirical density estimate for X at time t
$p_{X,t}^{\mathrm{cur}}(x)$	curiosity density for X at time t
$p_{X,t}^{\mathrm{tar}}[x \mid x(t)]$	target density for X at time t
$\Psi_X[x \mid x(t)]$	Gaussian modulator for target selection
$\mathcal{X}_X$	support/value space of var. X
$x^*(t)$	sampled target value at time t
$\hat{\boldsymbol{x}}(t)$	assignment over WM vars. at time t
$\mathcal{T}_{t:t+h'}^{\mathrm{sim}}$	simulated trajectory over WM vars.
h'	short planning horizon
$\boldsymbol{x}^{\mathrm{obs}}(t)$	assignment to observed WM vars. at time t
$\boldsymbol{x}^{\mathrm{lat}}(t)$	assignment to latent WM vars. at time t
$q_t\left(\boldsymbol{x}^{\mathrm{lat}}\right)$	approximate posterior over latent WM vars. at time t
$p_{\mathcal{M}^{\mathrm{wm}}}$	joint density induced by WM
Φ_t	forward query through WM at time t
$\hat{\boldsymbol{x}}^{\mathrm{tar}}$	predicted target-var. values
$\boldsymbol{u}$	residual assignment
$\bar{\boldsymbol{u}}$	expected residual assignment
$\boldsymbol{x}^{\mathrm{tar}}$	target-var. value vector
$p_t(\boldsymbol{x}^{\mathrm{tar}} \mid C_t)$	preference density over target-var. values at time t

C_t	current preference specification at time t		sensitivity of X_i to its parent X_j
$\mathcal{L}_t^{\text{act}}$	action-generation loss at time t	$p(\boldsymbol{x} \mid C_t)$	context-conditioned belief over WM states
$\tilde{\boldsymbol{a}}(t)$	candidate action at time t	$r_{j\to\text{tar}}^{C_t}$	target-query relevance of var. X_j under context C_t
$\boldsymbol{a}^*(t)$	optimized action at time t	Φ_t^{tar}	forward target query induced by current assignment targets
$Z_k^{(l+1)}(t)$	k-th abstraction var. at hierarchical level $l+1$ at time t	Σ_t	typed signature of artificial LoT at time t
S_k	support set of lower-level vars. for abstraction k	$\mathcal{O}_t$	set of Σ_t-represented objects
$h_k^{(l)}$	abstraction mechanism from lower-level vars. to abstraction var. $Z_k^{(l+1)}$	$\mathcal{P}_t$	set of grounded predicates in Σ_t
$\boldsymbol{x}_{S_k}^{(l)}(t:t+\Delta t)$	l-level var. values restricted to support set S_k over a time window	$\mathcal{A}_t$	set of action terms in Σ_t
		$\mathcal{R}_t$	set of typed relations among represented vars. in Σ_t
C_t	current context for WM contextualization	g_P	grounding map for predicate P
$\mathcal{G}_{\mathcal{M}^{\text{wm}}}^{C_t}(t)$	contextualized active subgraph of WM	S_P	support set of vars. grounding predicate P
$\boldsymbol{x}_{\text{pa}(X_i)}^{C_t}$	parent assignment induced by context C_t	$\boldsymbol{x}_{S_P}(t)$	current values of vars. selected by S_P
$r_{j\to i}^{C_t}$	local context-dependent sensitivity of var. X_i to its parent X_j	$[\![P(o_1,\ldots,o_n)]\!]_t$	probabilistic valuation of predicate $P(o_1,\ldots,o_n)$ at time t
$R_{j\to i}^{C_t}$	context-averaged		

Introduction

Cognitive systems do not encounter the world as isolated observers receiving already structured descriptions of objects, actions, and causal relations. In human cognitive development, the earliest forms of cognition emerge through the coupling between bodily action, sensory change, and the gradual stabilization of regularities across repeated interaction (Piaget, 1952; Varela et al., 1991; O'Regan & Noë, 2001; Noë, 2004). Before language and explicit symbolic reasoning become available, infants acquire increasingly structured knowledge and expectations about their own bodies and their surroundings through movement, touch, vision, other sensory modalities, and intervention-like exploration (Gopnik et al., 1999; Rochat, 1998; DiMercurio et al., 2018). This developmental perspective motivates the view that perception and action should not be modeled as independent modules connected only at the level of input and output. Rather, they form a perception-action loop in which perception provides action-relevant structure, action changes the conditions of perception, and internal models are progressively constructed from this coupling (Noë, 2004; Fuster, 2004; Engel et al., 2013).

The present thesis adopts this enactivist view as a computational design principle. The central problem we study here is how an embodied artificial agent could transform high-dimensional, noisy sensorimotor data into structured internal representations that support prediction, action generation, and, subsequently, other higher-level cognitive functions, and how to facilitate such operations on the resulting representations. We intentionally focus on pre-verbal and low-level human developmental stages as the minimal biological parallels we can model to achieve this set of cognitive representations and functions. Instead of assuming a mature symbolic knowledge base, linguistic supervision, or a task-specific reward function, the thesis investigates how object-centric perceptual representations and sensorimotor action-effect regularities can be computationally learned, represented, and operated on within a common architecture. This emphasis follows the broader symbol-grounding motivation: if later concepts, symbols, or linguistic expressions are to be meaningful for an agent, they must be connected to perceptual and action-based experience (Harnad, 1990; Barsalou, 1999; Vogt, 2002).

Two representational problems are especially important in this setting. First, we focus on the visual modality as one of the most informative modalities, both cognitively and computationally, especially in the embodied setting. Computationally, visual input must be transformed from raw observations into structured descriptions of objects,

object-local evidence, motion, appearance, and scene state. Such processing is essential as high-level reasoning cannot efficiently operate directly with the full continuous visual stream when the relevant regularities to observe and learn concern persistent, discrete entities and their relations (Marr, 2010). Visual perception is therefore treated here as an active and inferential process that provides object-centered, structured descriptions of what is happening in the environment (Kersten et al., 2004; Yuille & Kersten, 2006). Second, action-effect learning must distinguish structured causal or intervention-like relations from mere statistical association. Biologically, an embodied agent not only observes that variables co-vary; it acts, changes its own sensorimotor state, and can, in principle, use these changes to learn mechanisms that support prediction and action selection (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021).

The thesis addresses these problems by theoretically proposing a minimal neurosymbolic perception-action architecture for embodied sensorimotor learning. The architecture combines an object-centric visual perceptual subsystem with a structured sensorimotor world model. The perceptual subsystem is intended to map raw camera input into structured, interpretable representations that describe the observed scene, the objects in it, visual information, and the kinematic behavior of those objects. The world model's purpose is then to represent the variables extracted via the perceptual process together with the agent's proprioception, learn causal and correlational relationships between them, and use these relationships for prediction, target-conditioned action inference, and internally guided exploration. In this sense, the proposed system is neurosymbolic, as it combines learned continuous mechanisms with explicit variables, graph structure, object records, conceptual representations, and interpretable queries.

The aim of this work is not to build a complete autonomous cognitive architecture, but to provide a theoretical design and partial proof-of-concept implementation of its selected components. The primary contribution of our work is therefore theoretical, and the implementation is used to test whether the proposed representational decomposition can be instantiated computationally. This thesis does not aim to benchmark general computer vision, unrestricted causal discovery, full active inference, or robotic control. The thesis is organized around three aims, stated more precisely in Section 3.1:

1. to design an object-centric perceptual system that transforms visual input into structured conceptual and scene-level representations;
2. to design a sensorimotor world model that represents and learns action-effect regularities usable for prediction and action inference;
3. to formulate both components within a common perception-action architecture, so that perception, world modeling, and action generation can be treated as coupled processes.

These aims are evaluated through restricted experiments: the perceptual system is tested on recorded real-world object interactions, and a minimal differentiable causal

model is trained on proprioceptive motor-babbling transitions originating in robotic simulations. Hence, the experiments assess representational feasibility and compatibility rather than state-of-the-art robustness or autonomous performance.

The remainder of the thesis proceeds as follows. Chapters 1 and 2 introduce the theoretical background of multiple areas required for this project, and related work. Chapter 3 states the scope, requirements, and formal problem setting, while Chapter 4 presents the proposed architecture in theoretical detail. Finally, Chapter 5 reports the proof-of-concept implementations and experiments, while Chapter 6 interprets their implications and limitations and proposes several cognitively motivated, purely hypothetical extensions to the architecture.

Chapter 1

Preliminaries

This chapter provides a theoretical overview of the primary methods and approaches employed throughout this thesis. Here, we first introduce sensorimotor cognition as the embodied coupling of perception and action, and then outline visual perception as the process by which sensory input is transformed into structured, action-relevant representations. The chapter afterward presents learning and reasoning formalisms used later for modeling uncertainty, causal structure, sequential action generation, and multilevel neurosymbolic organization. Together, these preliminaries establish the conceptual basis we will refer to describe the proposed perception-action architecture and its computational components in the following chapters.

1.1 Sensorimotor Cognition

Sensorimotor cognition is grounded in the continuous coupling between perception and action, in which an agent learns, represents, and uses regularities among its movements, bodily states, and sensory consequences to understand and act in the world (Varela et al., 1991; O'Regan & Noë, 2001; Noë, 2004). As it biologically serves as an early-developing level of cognition, sensorimotor cognition provides a developmental basis for later perceptual, conceptual, and action-oriented capacities (Piaget, 1952; Varela et al., 1991; Noë, 2004; Gallagher, 2017). Following this definition, it necessitates an agent's embodiment, which translates to the need for robotic approaches when attempting to computationally model such a cognition.

In this work, we focus primarily on the sensorimotor domain in robotics and model a subset of sensorimotor cognitive functions of pre-verbal children, which we then partially integrate with visual perception,[1] and attempt to derive higher-level cognitive functions from this fusion. This section thus describes cognitive correlates of low-level sensorimotor cognition, learning, and representations, and it further maps these principles to their computational analogs.

[1] For an overview of cognitive and computational background of visual processing, see Section 1.2.

1.1.1 Cognitive Perspective

Sensorimotor cognition constitutes a foundational level of human cognitive development (Piaget, 1952) and is the first to develop in infants, beginning prenatally (Fagard et al., 2018; Myowa-Yamakoshi & Takeshita, 2006). Developmentally, sensorimotor skills precede linguistic and explicit symbolic reasoning: before infants can describe objects or relations, they learn regularities in movement, proprioception, touch, and changing sensory input (Gopnik et al., 1999).

Early Sensorimotor Organization

Infants enter postnatal development with early sensorimotor organization shaped by prenatal movement and bodily constraints (Fagard et al., 2018; Kanazawa et al., 2022; Adolph & Hoch, 2019), as well as with innate expectations (interpretable as priors in a Bayesian sense) about physical objects and events (Spelke & Kinzler, 2006; Spelke, 1990; Baillargeon, 2004). Subsequently, through the process of *circular reactions* (CRs; Piaget, 1952), infants build on and refine this organization by learning action-effect regularities pertinent to their bodies (i.e., primary CRs) first, expanding to the physical world later on (i.e., secondary CRs) (Augustyn et al., 2009). In accordance with developmental constructivism (Piaget, 1952), learning and expansion of causal knowledge is progressive and hierarchical. During the secondary CRs stage "infants begin to use causal behaviors [originating from the primary CRs] to recreate accidentally discovered, interesting effects" Augustyn et al., 2009, p. 33. CRs can thus be interpreted as a hierarchized intrinsic motivation for active physical exploration.

Concrete low-level mechanisms supporting this early sensorimotor organization include self-touch and spontaneous exploratory movement. Self-touch couples tactile, proprioceptive, and motor signals through the infant's own bodily activity, thereby contributing to bodily self-perception and body representation learning (Rochat, 1998; DiMercurio et al., 2018; Fagard et al., 2018; Gama, 2026). More generally, spontaneous exploratory movements provide variable action-sensory data from which action-effect regularities can be learned; in developmental robotics, this principle is commonly operationalized as motor babbling, as discussed in Section 1.1.2.

Sensorimotor Contingencies and Internal Models

The action-sensory relationships learned by infants in primary and secondary CRs can be characterized as *sensorimotor contingencies* (SMCs; O'Regan and Noë, 2001; Noë, 2004), i.e., "law-like relations between actions and contingent changes in the sensory signals" Maye and Engel, 2012, p. 106. While these relationships are initially discovered as weak associations (or correlations in statistical formalism) accidentally elicited by infants' own actions (Rochat, 1998; DiMercurio et al., 2018), they are solidified with accumulating observations. The product of this process corresponds to

approximate knowledge of causal relationships between children's own bodies and the surrounding environment. Such a level of *egocentric* (Woodward, 2011; Gärdenfors, 2006) causal knowledge and learning has been formally categorized as *Grade 1: Individual causal understanding* in the graded model of human causal cognition evolution by Lombard and Gärdenfors (2017; 2018) as well as *Category 1: Sensory-motor learning* and *Category 2: Learning about how the robot affects the world* of the analogous graded model of robotic causal cognition by Hellström (2021).

Learned SMCs can therefore be understood as rudimentary *internal models* (Miall & Wolpert, 1996; Wolpert & Kawato, 1998; Wolpert & Flanagan, 2001): they allow children to predict the sensory consequences of actions and to distinguish controllable from coincidentally correlated environmental changes (de Lange et al., 2018; Dogge et al., 2019). While during initial developmental stages internal models serve as predictors of motor action contingencies,[2] during later development these predictive structures become integrated into broader *intuitive theories* (Gerstenberg & Tenenbaum, 2017) of objects, agents, and physical and social relations beyond the purely sensorimotor domain (Lake et al., 2016; Goodman et al., 2011).

Active Learning and Embodiment

From the second year of life onward (Augustyn et al., 2009), children engage in more directed exploration by experimentation (Gopnik & Wellman, 1994; Gopnik et al., 1999; Gopnik & Schulz, 2004), resembling active Bayesian learning and inference[3] (Sobel et al., 2004; Schulz et al., 2007; Kayhan et al., 2019). In this context, in addition to ongoing sensorimotor learning through passive observation and accidental discovery, children begin to actively generate (predominantly sensorimotor) hypotheses about the world, test them by conducting *causal interventions* (Pearl, 2009), and revise them based on new ground-truth observations. By this stage, early learning is active, hierarchical, model-based, and intervention-sensitive. Furthermore, it ceases to be purely egocentric as children also begin causal learning by observing other agents' interventions and reproducing actions that yield similar effects (Meltzoff et al., 2012; Waismeyer & Meltzoff, 2017), thus expanding into the social domain.

This developmental trajectory, naturally, motivates a broader *embodied* and *enactivist* view of cognition (Varela et al., 1991; Gallagher, 2017). Perception is not a mere passive reconstruction of the external world but rather both forms and is formed by potential actions the agent can take within the environment, thereby actualizing the bidirectional co-dependence of perception and action (Noë, 2004). Hence, cognition can be understood as organized around a *perception-action loop* (PAL, Fig. 1.1; Neisser, 1976; Fuster, 2004), in which cognitive functions emerge from the continuous coupling

[2]This delineation is in accord with the original concept of internal models of motor control (Sperry, 1950; von Holst & Mittelstaedt, 1950).

[3]For an overview of the formal background of Bayesian learning, see Section 1.3.1.

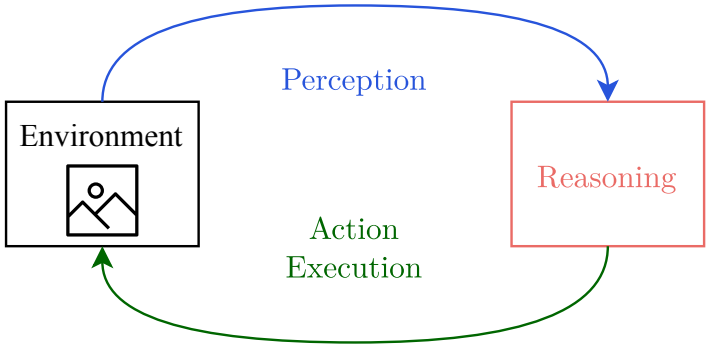


Figure 1.1: High-level schema of a generic perception-action loop. An agent receives a perceptual input from the environment, performs reasoning based on the input and its internal state, and produces an action applied to the environment.

between perception, action, and environmental feedback (Hommel et al., 2001; Engel et al., 2013).

To summarize, human cognition develops progressively and hierarchically from sensorimotor interaction toward more abstract forms of representation and reasoning. Infants first learn regularities between their actions and sensory observations through self-generated movement, accidental discovery, and repetition; later, exploration becomes increasingly directed, intervention-sensitive, and socially informed. The resulting action-sensory regularities elevated into causal relationships can be interpreted as internal models that allow children to predict and simulate environmental behavior without immediate action execution.

In this thesis, we use these cognitive principles to facilitate autonomous sensori- and visuomotor learning in humanoid robots embedded in simulated environments, as it can be seen that embodiment is crucial for the development of rich and robust cognitive functions and representations. Similar to humans, at a high level we base our cognitive architecture (Chapters 3 and 4) on a modeled PAL and derive other, more complex cognitive functions from this framework (Section 6.4). For an overview of generic computational approaches to some of the concepts described in this section, refer to Section 1.1.2 as well as to Section 2.1 for the discussion on the state of the art.

1.1.2 Computational Perspective

The cognitive account outlined in Section 1.1.1 motivates a computational approach, in which cognitive functions are not modeled as mature, fully developed mechanisms in isolation from an environment and without complex integration with other cognitive functions and representations. Instead, they progressively emerge from embodied sensorimotor (at minimum) interaction. This approach is well-realized in the field of developmental robotics (Cangelosi & Schlesinger, 2015), where the construction of embodied agents and their embedding in real or simulated environments that learn through action allow an active development of cognitively inspired perception and action mechanisms (Asada et al., 2009).

This contrasts with much recent work on large language, multimodal vision-language, and vision-language-action models, for instance, which, while impressive in task-level performance, do not by themselves explain how cognitive functions become grounded, robust, and generalizable through embodied interaction. Such models can learn relatively robust statistical regularities from large-scale observational data, but this does not guarantee (Lake, 2014; Farkaš et al., 2025) the grounded causal world models (Harnad, 1990; Vogt, 2002; Lake et al., 2016) needed for systematic out-of-distribution generalization (Schölkopf et al., 2021), reliable perception-action coupling, or resistance to hallucination-like failures (Huang et al., 2023; Bai et al., 2024; Favero et al., 2024).

In the sensorimotor domain of developmental robotics, the initial problem is inducing a predominantly uncalibrated agent to acquire informative sensorimotor regularities. Analogously to the early exploratory mechanisms of infants described in Section 1.1.1, artificial agents use self-generated movement to collect action-sensory data and learn predictive relations between, for example, motor commands, proprioceptive states, tactile feedback, and visual observations. This mechanism is commonly operationalized as *motor babbling*, where the agent explores its action space by producing variable, usually randomized, motor actions and observing their sensory consequences (Saegusa et al., 2009; Cibula et al., 2024; Gama, 2026), or, more recently, as its more sophisticated version, *goal babbling* (Rolf et al., 2010; Baranes & Oudeyer, 2013; Gama et al., 2023), which biases exploration toward reachable or informative goal states. Goal babbling and more advanced explorative approaches are often guided by *intrinsic motivation*, which provides internal learning signals (Oudeyer & Kaplan, 2007) in the absence of or in addition to extrinsic task rewards (Sutton & Barto, 2018). An intrinsic signal may mimic curiosity in living agents (Schmidhuber, 2010), and therefore drive an artificial agent to seek novelty, prediction-error reduction, or information gain (Schillaci et al., 2016; Aubret et al., 2023; Salge et al., 2014).

The action-sensory relationships learned through such exploration can be formalized as computational *forward* (FM) and *inverse models* (IM) (Wolpert & Kawato, 1998; Nguyen-Tuong & Peters, 2011), which are analogous to biological internal mod-

els discussed in Section 1.1.1, and in the same fashion can be used to facilitate body schema emergence (Nguyen et al., 2021; Gama, 2026), sensorimotor prediction (Cibula, 2024), and can be scaled up to more abstract domains (Gerstenberg & Tenenbaum, 2017; Lake et al., 2016; Goodman et al., 2011; Hellström, 2021). In the general computational context (Francis & Wonham, 1976; Cibula et al., 2024), FMs (Dearden & Demiris, 2005) predict the ground-truth sensory or, more formally, state consequences $\boldsymbol{s}(t + \Delta t) \in \mathcal{S}$ of the agent's action $\boldsymbol{a}(t) \in \mathcal{A}$ executed in $\boldsymbol{s}(t) \in \mathcal{S}$, i.e., acting as a predictor

$$\text{FM}\colon\ [\boldsymbol{s}(t), \boldsymbol{a}(t)] \mapsto \hat{\boldsymbol{s}}(t + \Delta t), \tag{1.1}$$

where $\hat{\boldsymbol{s}}(t + \Delta t)$ denotes a predicted state $\Delta t \in \mathbb{N}^+$ discrete time steps after the execution of $\boldsymbol{a}(t)$, and $\mathcal{S}, \mathcal{A}$ are abstract state and action spaces, respectively. Complementarily, IMs estimate an action $\boldsymbol{a}(t)$ to perform in $\boldsymbol{s}(t)$, required to achieve a desired future state $\boldsymbol{s}(t + \Delta t)$:

$$\text{IM}\colon\ [\boldsymbol{s}(t), \boldsymbol{s}(t + \Delta t)] \mapsto \hat{\boldsymbol{a}}(t). \tag{1.2}$$

Formally, many of these mechanisms are implemented within reinforcement learning (RL) frameworks, where the agent interacts with the environment formalized as a Markov decision process (MDP), and learns policies from reward signals generated by the task (Sutton & Barto, 2018). However, as discussed above, in developmental settings, the reward signal is often intrinsic and epistemic rather than purely task-defined. The present thesis adopts this interaction-centered view of agent–environment coupling, as it aligns with the enactivist principles we aim to follow; however, we do not treat reward maximization as the primary explanatory principle (see Introduction and Chapter 3 for more details).

Once body-centered contingencies are established, the same interaction-centered logic can be extended to the external environment. Through physical interaction, embodied agents can learn how their actions affect objects (Fitzpatrick et al., 2003; Fitzpatrick, 2003; Lohmann et al., 2020; Kruzliak et al., 2024), reveal affordances (Gibson, 2014), and produce predictable environmental changes (refer to Section 1.2.2 for more details). This shifts sensorimotor learning from bodily calibration toward larger-scale, object- and action-centered world modeling, in which world entities are represented by the effects they afford under intervention.

For a more comprehensive overview of developmental robotics, and specifically sensorimotor learning, exploratory strategies, and body representations, refer to Gama (2026), Zheng et al. (2025), Nguyen et al. (2021), and Schillaci et al. (2016). These mechanisms motivate the world-model component of the proposed architecture, where sensorimotor regularities are represented as structured variables and mechanisms usable for prediction and action inference.

1.2 Visual Perception

Visual perception provides an embodied agent with structured information about its surrounding environment by transforming high-dimensional, often noisy, sensory input into representations of objects, spatial relations, events, and action-relevant properties (Marr, 2010; Kersten et al., 2004; Yuille & Kersten, 2006; Gibson, 2014). In biological cognition, vision is not a passive reconstruction of the external world, but an active and interpretive process shaped by attention, prior knowledge, embodiment, and ongoing action (O'Regan & Noë, 2001; Noë, 2004; Aloimonos et al., 1988; Clark, 2013; Démuth, 2013; de Lange et al., 2018). In this work, visual perception is treated as a computational and cognitive bridge between low-level sensorimotor experience and higher-level world modeling, providing perceptual representations that can later be integrated with sensorimotor and causal learning.

1.2.1 Neurocognitive Correlates

Organization of the Human Visual System

The human visual system is hierarchically organized, progressively transforming retinal stimulation into increasingly abstract representations of features, objects, spatial relations, and scenes (Hubel & Wiesel, 1962; Rolls, 2016; O'Reilly et al., 2024). The visual pathway (Fig. 1.2) begins with the transduction of the visual stimuli into neural signals by photoreceptive cells in the retinas and retinal ganglion cells (RGCs). The signal then propagates down the optic nerves, crosses at the optic chiasm, and continues via the optic tract to the lateral geniculate nuclei (LGN), which are dedicated early-processing areas in the thalamus. Subsequently, the signal continues via the geniculostriatal pathway (i.e., the optic radiation), terminating in the *primary visual cortex* (V1) located in the occipital lobe (O'Reilly et al., 2024). Early visual processing in the RGCs, LGN, and V1 entails the progressive detection of luminance contrast, center-surround structure, and, finally, oriented edges (Hubel & Wiesel, 1962).

From V1 onward, visual processing splits into two streams with broader functional specialization (Goodale & Milner, 1992; Ungerleider & Mishkin, 1982): the *ventral* (VVP) and *dorsal pathway* (DVP) (Fig. 1.3). Specifically:

> The projections going in a ventral direction from V1 to V4 to areas of inferotemporal cortex [ITC] [...] are important for recognizing the identity ("what") of objects in the visual input, while those going up through parietal cortex extract spatial ("where") information, including motion signals in area MT [middle temporal visual area] and MST [medial superior temporal area]. O'Reilly et al., 2024, p. 111

Functionally, the VVP performs a progressive abstraction of visual information. Continuing from local oriented edges detected in V1, in V2 these are combined into junc-

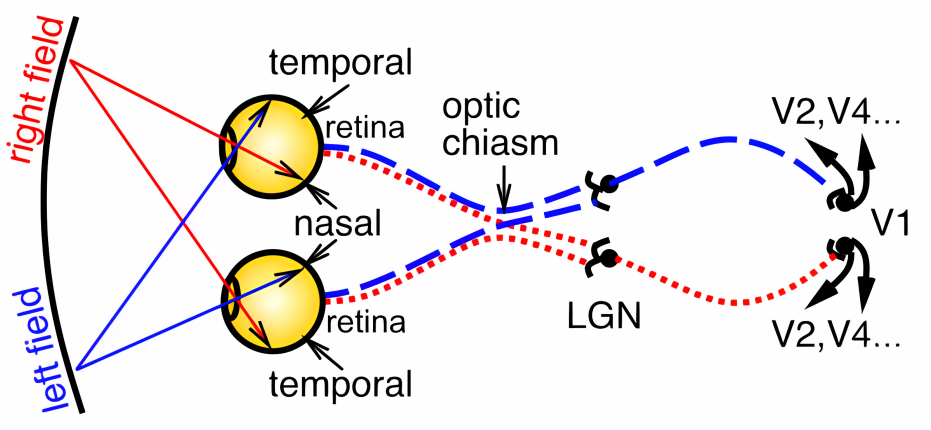


Figure 1.2: Early visual processing pathway propagating a visual stimulus detected by the retina to the V1 cortex in the occipital lobe (O'Reilly et al., 2024, Figure 6.1).

tions, textures, and other elementary shape features; in V4, progressively more complex shape features are represented over larger regions of the visual field. Further along the pathway, *posterior inferotemporal cortex* (PIT) represents whole object shapes with an increasing invariance to retinal position, scale, and orientation, while *anterior inferotemporal cortex* (AIT) supports highly abstract object representations approaching semantic content (O'Reilly et al., 2024). Thus, ventral processing can be understood as a transformation from local sensory structure to stable object-centered representations suitable for recognition (Rolls, 2016; Ayzenberg & Behrmann, 2024; von der Heydt, 2023).

Meanwhile, the DVP exhibits a complementary functional progression. Beginning from V1, the pathway carries information through motion-sensitive regions (i.e., MT and MST) toward the parietal cortex, where visual information is increasingly represented in terms of spatial location, motion, depth, and action-relevant relations between the agent's body and the environment (O'Reilly et al., 2024; Goodale & Milner, 1992). Dorsal processing is therefore primarily concerned with locating objects, tracking their movement, and transforming visual information into spatial representations useful for attention and action.

For a more comprehensive overview of the functional neuroanatomy of the human visual system, as well as other neurocognitive correlates of visual perception, refer to O'Reilly et al. (2024, Chapter 6).

Bidirectional Processing

Despite the hierarchical organization described above, visual processing is generally accepted to be a bidirectional, bottom-up/top-down process (Dijkstra et al., 2017; McMains & Kastner, 2011; Mumford, 1992). Bottom-up sensory evidence interacts continuously with top-down influences from higher cognitive processes, including attention (Desimone & Duncan, 1995; Deutsch & Deutsch, 1963; Cowan, 1998), prior

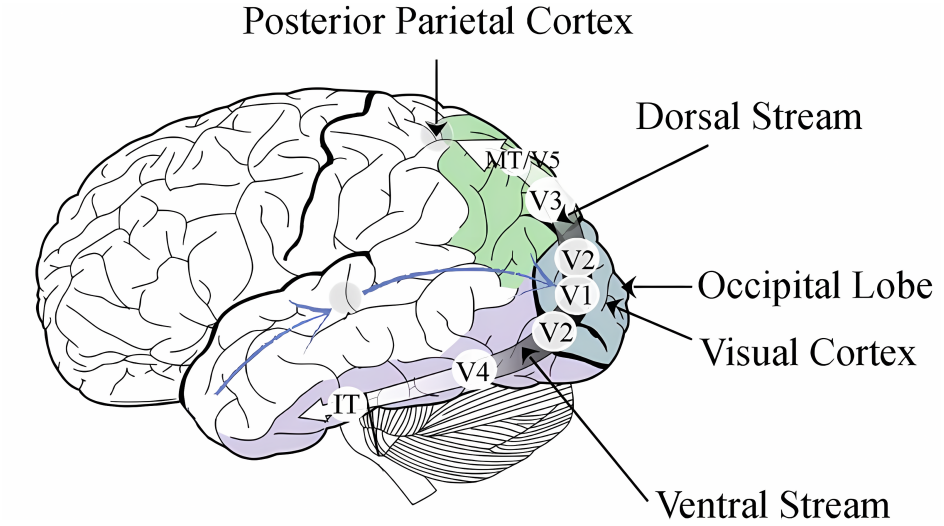


Figure 1.3: Neural anatomy of the ventral and dorsal visual pathways starting in the V1 cortex and projecting to the temporal and parietal lobes, respectively. Adapted from Zareh et al. (2024).

knowledge (Teufel et al., 2018; Adams et al., 2004), environmental context (Bar, 2004; Oliva & Torralba, 2007), and expectations[4] (Summerfield & de Lange, 2014; Peelen et al., 2024). Visual perception is therefore better understood as a recurrent process, in which information propagates both upward from sensory input and downward from more abstract representations (Spalek et al., 2024). Predictive-coding accounts formalize this bidirectionality by interpreting perception as inference, in which higher levels generate predictions and lower levels communicate prediction errors when sensory evidence deviates from them (Rao & Ballard, 1999; Bastos et al., 2012; de Lange et al., 2018; Peelen et al., 2024). For an overview of the formal framework of predictive coding, see Section 1.3.3.

Semantic Representations

Perceptual processes (especially in the visual system) contribute to higher-level representations of percepts in a form that can be retained, generalized, and used beyond the immediate sensory episode. A *concept* may be generally understood as a cognitive encoding of a class of entities, properties, relations, or events (Medin & Coley, 1998; Gärdenfors, 2000), while a *symbol* is a discrete representational unit that can stand for such content in reasoning, for instance (DeLoache, 2004; Harnad, 1990; Barsalou, 1999). In the visual domain, semantic cognition is thus fundamentally object-centric, as non-symbolic sensory input is organized into persistent entities that can be recognized

[4]Note that, in this context, we differentiate between expectations understood as Bayesian priors predominantly in an active inference context, and prior knowledge as previously learned structure. While prior knowledge generates expectations in a Bayesian sense, expectations may also be a product of other factors.

as the same objects across changing contexts and associated with their properties, relations, and possible behaviors.

Semantic knowledge is commonly not understood as being stored in amodal symbols (Barsalou, 1999), but is distributed across perceptual and action modalities associated with a concept's relevant properties (Lambon Ralph et al., 2016; Takáč, 2007; Farkaš et al., 2012). For concrete concepts in the visuomotor domain, semantic content may therefore include visual features, action-related properties, and other modality-specific features. This distributed organization is consistent with the view that semantic representations are grounded in perception and action (Barsalou, 1999; Newell et al., 2023).

At the neural level, semantic cognition requires integration across modality-specific areas (Patterson et al., 2007). The *anterior temporal lobe* (ATL) is interpreted as a convergence region (*semantic hub*), supporting increasingly abstract and multimodal conceptual representations by integrating representations in the modality-specific areas (*modal spokes*) (Patterson et al., 2007; Lambon Ralph et al., 2016; Frisby et al., 2023; O'Reilly et al., 2024). In this sense, the anterior-most stages of the VVP transition increasingly invariant visual object representations toward broader, more abstract semantic knowledge, as anterior temporal regions integrate visual information with other modalities (Frisby et al., 2023).

For a further overview of cognitive semantic representations from the neural and computational point of view, see Takáč (2007), Lambon Ralph et al. (2016), and Frisby et al. (2023).

1.2.2 Active Computer Vision

Active computer vision (ACV) departs from the traditional computational paradigm that visual perception operates on a fixed, passively received image stream. Instead, an agent can actively improve its perception by controlling its sensors, selecting viewpoints, directing attention, and generating informative observations through movement or interaction (Bajcsy, 1988; Aloimonos et al., 1988). Therefore, in alignment with ecological accounts of sensorimotor cognition and perception (see Sections 1.1.2 and 1.2.1, respectively), ACV provides a computational analog to bidirectional visual processing (O'Regan & Noë, 2001; Noë, 2004), and, as such, fits within a PAL (Mirza et al., 2016; Fuster, 2004).

ACV is especially relevant for object-centered vision, since objects in the scene are not given directly in the sensory stream as already separated units, but must be individuated from changing visual input and represented as persistent entities despite motion, occlusion, and viewpoint change. Developmental research suggests that cues such as cohesion, continuity, and common motion support this process (Spelke, 1990; Baillargeon, 2004). Active self-generated movement and physical interaction can expose these cues by, for example, causing object parts to move together, separating manipulated objects from the background, and revealing changes in pose, articulation, and

interaction-dependent physical behavior that are not available from a static image alone (Held & Hein, 1963; Fitzpatrick et al., 2003; Lohmann et al., 2020). ACV methods can thus provide generally richer and more robust scene- and object-level information that can be leveraged for various computer vision tasks, such as instance segmentation (Fitzpatrick, 2003; Katz et al., 2013) or object kinematic modeling (Nematollahi et al., 2020; Kruzliak et al., 2024).

In this context, a related notion of *affordances* (Gibson, 2014) should be mentioned, i.e., "the action possibilities offered to an animal [an agent] by the environment with reference to the animal's action capabilities" Osiurak et al., 2017, p. 403. From this perspective, object representations are characterized not only by their appearance but also by the actions they enable and the consequences those actions produce. In computational terms, affordance learning has motivated approaches that associate object perception with action understanding (Gupta & Davis, 2007), learn object properties through interaction (Lohmann et al., 2020), or discover actionable object parts and their possible movements through interaction (Gadre et al., 2021).

1.2.3 Conceptual Spaces

As discussed in Section 1.2.1, semantic representations organize non-symbolic perceptual experience into more stable conceptual structure. This raises a central aspect of the symbol grounding problem: if symbols are supposed to be "meaningful," they must ultimately be connected (i.e., *grounded*) to non-symbolic (i.e., subsymbolic, by implication) perceptual and action-based experience (Harnad, 1990; Vogt, 2002). Since perceptual input, for animals and machines alike, is continuous, high-dimensional, and similarity-structured (Douven et al., 2023), a suitable semantic representation should preserve similarities between observations, while supporting categorization (Barsalou, 1999). Such a representation must therefore organize raw perceptual input into higher-level entities while preserving the similarity relations from which categorization can emerge.

One of the frameworks satisfying these requirements is the theory of *conceptual spaces* (Gärdenfors, 2000, 2014; Gärdenfors & Osta-Vélez, 2023). By its definition (Gärdenfors, 2000), a conceptual space is a geometric space that

> consists of a number of *quality dimensions* such as color, pitch, temperature, weight, and the three ordinary spatial dimensions. The quality dimensions are endowed with certain topological or metric structures; some quality dimensions can have a discrete structure. [...] Dimensions of a conceptual space correspond to attributes of represented objects. Because not all attributes are relevant to all represented entities, dimensions are organized in *domains*. Takáč, 2007, p. 36

For our purposes, we use the following formalization. Let $D = \{d_1, \ldots, d_n\}$ denote

the quality dimensions available to our representational system, so the full conceptual space can be defined as the Cartesian product

$$\mathcal{C} = \prod_{d \in D} d. \tag{1.3}$$

Following Gärdenfors (2000), dimensions within the same domain are labeled *integral* as assigning a value to one requires assigning values to the others as well. By contrast, *separable* dimensions belong to different domains and can be evaluated independently (Coninx, 2022). Therefore, a domain can be represented as

$$\delta_j = \prod_{d \in I_j} d, \tag{1.4}$$

where a dimension set $I_j \subseteq D$ specifies δ_j-relevant integral quality dimensions, with $\delta_j \in \mathcal{D}$ while $\mathcal{D} = \{\delta_1, \ldots, \delta_m\}$ represents all the domains within $\mathcal{C}$. Note that the sets $I_1, \ldots, I_m$ are pairwise disjoint. As not all domains are relevant to every concept or conceptual comparison, by

$$\mathcal{C}_J = \prod_{\delta \in J} \delta \tag{1.5}$$

we denote the conceptual subspace induced by the relevant subset of domains $J \subseteq \mathcal{D}$.

An observed entity represented in this subspace is a point $\boldsymbol{x} \in \mathcal{C}_J$, and we use $\boldsymbol{x}_\delta \subseteq \boldsymbol{x}$ to denote its subvector restricted to the dimensions of the domain δ. Similar to any metric space, the semantic similarity between two represented entities $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}_J$ can be modeled by some metric dist, such as

$$\mathrm{dist}_{\mathcal{C}_J}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sum_{\delta \in J} w_\delta \, \mathrm{dist}_\delta(\boldsymbol{x}_\delta, \boldsymbol{y}_\delta), \tag{1.6}$$

where dist_δ denotes a within-domain metric of the domain δ, and $w_\delta > 0$ is the context-dependent salience weight of that domain, while it must hold that $\sum_{\delta \in J} w_\delta = 1$. In the standard metric treatment of conceptual spaces, integral dimensions within a domain are typically compared by a weighted Euclidean distance (i.e., a weighted L_2 metric), whereas separable domains are combined additively by a weighted Manhattan distance (i.e., a weighted L_1 metric) (Gärdenfors, 2000; Bechberger & Kühnberger, 2017). Smaller distances correspond to greater similarity.

Within this formalism, observed instances $\boldsymbol{x} \in \mathcal{C}_J$ correspond to *exemplars* (Medin & Schaffer, 1978; Nosofsky, 1984). A natural property is represented as a convex region within a single domain (Gärdenfors, 2000, 2014). For instance, "red" is represented as a convex region in the color domain. A concept such as "apple", by contrast, combines regions from several domains, such as color, shape, size, and taste, together with the relative salience of those domains and possible correlations between them (Gärdenfors, 2000; Gärdenfors & Osta-Vélez, 2023). Repeatedly observed similar exemplars occupy nearby locations in the space, and their central tendency $\boldsymbol{\mu}_\ell \in \mathcal{C}_J$ can thus be understood as a *prototype* (Rosch, 1978).

If category membership is operationalized by nearest-prototype assignment, the prototypes $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ induce a Voronoi tessellation of the relevant conceptual subspace (Gärdenfors & Williams, 2001):

$$V_\ell = \{\boldsymbol{x} \in \mathcal{C}_J \mid \mathrm{dist}_{\mathcal{C}_J}(\boldsymbol{x}, \boldsymbol{\mu}_\ell) \leq \mathrm{dist}_{\mathcal{C}_J}(\boldsymbol{x}, \boldsymbol{\mu}_r), \forall r \neq \ell\}, \tag{1.7}$$

where V_ℓ denotes the region assigned to category ℓ. In other words, categorization in this sense can be understood as partitioning a continuous perceptual space according to similarity to stored prototypes.

In our object-centric context, the multidomain nature of conceptual spaces is useful, as it enables the representation of objects beyond the visual domain alone. The perceptual semantics in our case[5] may also include spatial and topological properties, as well as kinematic and event-related properties, such as motion and interaction (Mandler, 1992; Mandler, 2007; Gärdenfors, 2019, 2020, 2024). Conceptual spaces can therefore represent object concepts through coordinated visual, spatial, and behavioral domains, which is especially relevant when objects are learned through active vision (see Section 1.2.2).

Finally, it should be noted that although the present formalization treats conceptual spaces metrically, they fundamentally differ from standard distributional semantic vector spaces (Mikolov et al., 2013; Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019), whose vectors are learned primarily from statistical regularities in linguistic data (Turney & Pantel, 2010). Such models, while powerful in representing linguistic similarity, do not generally operate with interpretable, disentangled quality dimensions and, in their standard textual-only form, lack complex perceptual grounding (Harnad, 1990; Roy, 2008).

For a more extensive overview of the theory of conceptual spaces as well as related methods of computational semantics, refer to Gärdenfors (2014) and Takáč (2007), respectively.

[5]Refer to Chapter 3 and Section 4.2 for more details on approaches we use in this domain.

1.3 Learning and Reasoning

In the computational context, learning and reasoning provide the formal mechanisms by which an embodied agent can update its internal representations and, with limited and imperfect information, infer unobserved states of the environment and select appropriate actions (Pearl, 1988; Murphy, 2012; Parr et al., 2022). While the previous sections described how an agent is coupled to the world through bodily interaction and sensory processing, the present section describes how such interaction can be organized into structured beliefs, causal knowledge, and goal-directed behavior. In this research, these mechanisms are approached primarily through probabilistic learning (Section 1.3.1), causal inference (Section 1.3.2), sequential decision-making, active inference (Section 1.3.3), and neurosymbolic multilevel modeling (Section 1.3.4), which together provide the theoretical background for the structured world models developed in Section 4.3.

1.3.1 Bayesian Learning and Probabilistic Graphical Models

Bayesian learning provides a computational formalism of how *beliefs* can be constructed and revised when new *evidence* is introduced (Pearl, 1988; Murphy, 2012). This is particularly useful when modeling complex environments from imperfect observations, as in our case of internally modeling an agent's surrounding environment from possibly imprecise and noisy visual and sensorimotor observations.

To introduce the formalism, let H and D denote random variables (RVs) over possible hypotheses and observations, respectively. For a particular hypothesis $H = h$ and observed data point $D = d$, the posterior probability of h after observing d is given by Bayes' rule:

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)}, \tag{1.8}$$

where $p(h)$ denotes the prior probability of the hypothesis, $p(d \mid h)$ the likelihood of observing d given h is true, and $p(d)$ the marginal probability of the observation. Learning in this sense, therefore, consists of updating prior beliefs based on how well competing hypotheses predict the available evidence (Murphy, 2012, Section 2.2).

Real-world problems, however, usually involve multiple uncertain variables related by conditional dependencies. A *Bayesian network* (BN; Pearl, 1985) is a directed probabilistic graphical model (PGM) that represents such a joint probability distribution as a directed acyclic graph (DAG) whose vertices represent modeled RVs and edges encode direct probabilistic dependencies between them (Pearl, 1988; Koller & Friedman, 2009).

For RVs X, Y, and Z, we write $X \perp\!\!\!\perp Y \mid Z$ to denote that X is conditionally independent of Y given Z, i.e.,

$$p(X = x \mid Y = y, Z = z) = p(X = x \mid Z = z), \tag{1.9}$$

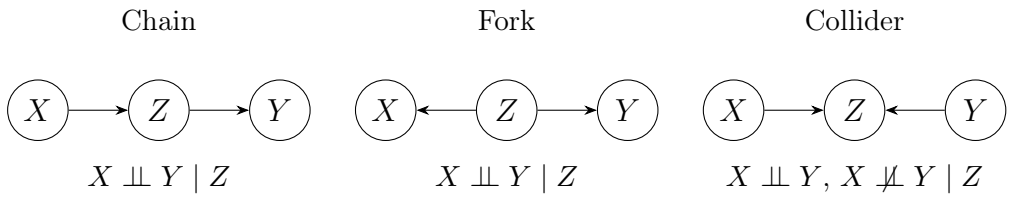


Figure 1.4: Three elementary topologies in BNs. In chains and forks, conditioning on the intermediate variable Z blocks dependence between X and Y; in colliders, observing Z can induce dependence between its parents.

whenever the corresponding conditional distributions are defined. Intuitively, once Z is known, additional knowledge of Y provides no further information about X. Beyond displaying direct probabilistic dependencies, the graphical structure of a BN also encodes conditional-independence assumptions among the modeled variables. Under the local Markov property, each variable is conditionally independent of its non-descendants given its parents (Pearl, 1988; Koller & Friedman, 2009).

The effect of graph structure on conditional independence can be illustrated by three elementary graph topologies (Fig. 1.4). In both a chain ($X \to Z \to Y$) and a fork ($X \leftarrow Z \to Y$), conditioning on the intermediate variable Z blocks the dependence between X and Y, such that $X \perp\!\!\!\perp Y \mid Z$. In a collider ($X \to Z \leftarrow Y$), by contrast, X and Y are marginally independent, but can become conditionally dependent once their common effect Z is observed. These topologies determine the conditional-independence relations encoded by more complex BNs (Pearl, 1988; Koller & Friedman, 2009).

Let us now consider a set of RVs $\mathcal{X} = \{X_1, \ldots, X_n\}$ with the corresponding joint value assignment $\boldsymbol{x} = [x_1, \ldots, x_n]$, and let $\mathrm{pa}(X_i)$ denote the parent vertices of X_i in the graph. We use $\boldsymbol{x}_{\mathrm{pa}(X_i)}$ to denote the value assignment to these parents induced by $\boldsymbol{x}$. Under the local Markov property, the joint distribution over $\mathcal{X}$ factorizes as

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p\left[x_i \mid \boldsymbol{x}_{\mathrm{pa}(X_i)}\right]. \tag{1.10}$$

This factorization thus allows the joint distribution to be specified through local conditional distributions rather than modeled directly as a single distribution (Pearl, 1988).

Because of these properties, BNs are suitable for modeling uncertain systems composed of several interdependent variables while preserving an explicit representation of their dependency structure. Given new observations of some variables, beliefs about unobserved variables can be inferred by conditioning the joint distribution on the available evidence. When additional data are obtained, the parameters of the network, or, in more complex cases, the graphical structure itself, may also be learned from data (Koller & Friedman, 2009; Murphy, 2012). Consequently, BNs provide a versatile formalism for probabilistic learning of structured world models and inference over them. It is important to note, however, that edges of standard BNs encode only probabilis-

tic dependence rather than causal influence. Causal interpretation requires additional assumptions and formalism (Pearl, 2009); see Section 1.3.2 for an overview.

1.3.2 Causal Inference and Autonomous Experimental Design

As mentioned in Section 1.3.1, the probabilistic dependencies encoded by BNs do not alone distinguish causal relationships from correlational ones. Two variables may covary because one directly affects the other (i.e., a proper causal relationship), because both share a common cause (i.e., a confounding variable), or because of other dependencies induced by the graph structure. For instance, the fork topology (Fig. 1.4) introduced in Section 1.3.1 (i.e., $X \leftarrow Z \rightarrow Y$) may produce an association between X and Y even when neither variable causally influences the other. Causal inference addresses this limitation by modeling how a system would dynamically respond to interventions, rather than only how its variables co-occur under passive observation (Pearl, 1995, 2009, 2019; Peters et al., 2017; Schölkopf et al., 2021).

Structural Causal Models

A standard formal framework for this purpose is the *structural causal model* (SCM; Pearl, 2009; Peters et al., 2017; Schölkopf, 2022). Considering a set of RVs $\mathcal{X} = \{X_1, \ldots, X_n, U_1, \ldots, U_n\}$, following the standard definition (Pearl, 2009; Peters et al., 2017; Toth et al., 2022; Busch et al., 2025), an SCM can be represented as

$$\mathcal{M} \triangleq \langle \mathcal{V}, \mathcal{U}, \mathcal{F}, p(\mathcal{U}) \rangle, \tag{1.11}$$

where $\mathcal{V} = \{X_1, \ldots, X_n\}$ is a set of *endogenous variables* determined within the model (i.e., usually observed variables within the system), $\mathcal{U} = \{U_1, \ldots, U_n\}$ denotes a set of *exogenous variables* determined outside the model (i.e., unobserved background or noise variables outside of the system), $\mathcal{F} = \{f_1, \ldots, f_n\}$ is a set of deterministic *structural mechanisms*, and $p(\mathcal{U})$ is a distribution over $\mathcal{U}$. In standard (simpler) cases, each endogenous variable $X_i \in \mathcal{V}$ is assigned by a structural equation

$$X_i \triangleq f_i[\mathrm{pa}\,(X_i)\,, U_i]\,, \tag{1.12}$$

with $\mathrm{pa}\,(X_i) \subseteq \mathcal{V} \setminus \{X_i\}$ denoting the direct endogenous causes (i.e., causal parents) of X_i, and $U_i \in \mathcal{U}$ denoting the exogenous variable potentially influencing the relationship. Given this definition, as Toth et al. (2022, p. 16,263) note, "[f]rom a Bayesian perspective, an SCM can be viewed as a hierarchical data-generating process involving latent [i.e., exogenous] random variables" and locally specified causal mechanisms. The causal mechanisms in $\mathcal{F}$ induce a directed causal graph $\mathcal{G}_{\mathcal{M}}$, whose vertices represent the endogenous variables in $\mathcal{V}$ and in which an edge $X_j \rightarrow X_i$ is present whenever $X_j \in \mathrm{pa}\,(X_i)$ (Peters et al., 2017; Toth et al., 2022). Thus, in contrast to standard BNs, the edges of $\mathcal{G}_{\mathcal{M}}$ are interpreted as direct causal relations specified by the causal

mechanisms of the SCM. The model induces an *observational distribution* $p(\mathcal{V} \mid \mathcal{M})$ over the endogenous variables, obtained from the mechanisms together with the distribution over exogenous variables $p(\mathcal{U})$ (Pearl, 2009; Toth et al., 2022). Throughout this work, we will assume the standard acyclic Markovian SCMs, in which $\mathcal{G}_{\mathcal{M}}$ is a DAG and the exogenous variables are mutually independent, so the induced observational distribution obeys a BN factorization (Eq. 1.10) over the associated causal graph (Tian & Pearl, 2000; Pearl, 2009).

Supported Causal Operations

The additional expressive power of SCMs appears under *interventions*, i.e., external alterations of their structural mechanisms $f_i \in \mathcal{F}$. An intervention setting an RV X_i to a value x is denoted by $\mathrm{do}(X_i = x)$ and is represented by replacing the original mechanism determining X_i with the constant assignment $X_i \triangleq x$, while leaving the remaining mechanisms unchanged (Pearl, 2009). The resulting *interventional distribution*, for instance $p[X_j = y \mid \mathrm{do}(X_i = x)]$, describes the probability of $X_j = y$ after actively setting X_i to x. It is therefore generally distinct from the observational conditional probability $p(X_j = y \mid X_i = x)$, which conditions only on *observing* $X_i = x$ in the unmodified system. Pearl's *do-calculus* (Pearl, 1995, 2009) provides graphical rules for transforming expressions containing the do-operator and, when the causal graph permits it, for identifying interventional distributions from observational quantities (Peters et al., 2017).

SCMs also support *counterfactual* queries, which examine what would have happened under an alternative intervention in a particular realized situation. Their evaluation proceeds by *abduction*, *action*, and *prediction*: inferring the exogenous conditions compatible with factual evidence, modifying the model by a hypothetical intervention, and predicting the resulting outcome (Pearl, 2009; Pearl et al., 2016; Busch et al., 2025). Observational, interventional, and counterfactual queries form the three levels of Pearl's *"Ladder of Causation"* – (1) *association*, (2) *intervention*, and (3) *counterfactuals*, respectively (Pearl, 2009, 2019; Pearl & Mackenzie, 2018; Busch et al., 2025).

Autonomous Experimental Design

Finally, because an agent within this framework can model and execute interventions in its environment and therefore does not need to rely solely on passive observational data, causal learning also becomes a problem of efficient experimental choice. *Bayesian experimental design* (BED; Lindley, 1956) implements a possible solution by considering a set of candidate experiments Ξ and selecting the optimal experiment $\xi^* \in \Xi$ whose possible outcomes are expected to be most "useful" for the agent (Chaloner & Verdinelli, 1995; Ryan et al., 2016). Formally,

$$\xi^* \triangleq \operatorname*{argmax}_{\xi \in \Xi} \mathbb{E}_{d \sim p(d|\xi)} \left[u(\xi, d) \right], \tag{1.13}$$

where d denotes a possible outcome of experiment ξ, $p(d \mid \xi)$ is the predictive distribution over its possible outcomes, and $u(\xi, d)$ the utility assigned to obtaining outcome d after performing ξ. When the utility is defined as information gain, for example, by the Kullback-Leibler divergence (Kullback & Leibler, 1951) from the posterior belief $p(h \mid \xi, d)$ to the prior one $p(h)$ over hypotheses,

$$u(\xi, d) = D_{\mathrm{KL}}\big[p(h \mid \xi, d) \,\big\|\, p(h)\big], \tag{1.14}$$

the preferred experiment is the one expected to maximally reduce uncertainty about the relevant hypotheses (Long et al., 2013).

In causal settings, candidate experiments are interventions, and the hypotheses may concern causal structure, structural mechanisms, or both. *Active Bayesian causal inference* (ABCI; Toth et al., 2022) combines causal learning mechanisms with intervention optimization, thereby selecting interventions expected to be maximally informative about the target causal model or query. This formalizes how an embodied agent may acquire more complex causal knowledge not only from passive observation, but also through autonomously chosen informative interventions.

1.3.3 Sequential Decision-Making and Active Inference

The previous sections introduced probabilistic modeling methods for learning and reasoning under uncertainty. As discussed in Section 1.1, however, in an agentic context, beliefs are updated through time as the agent perceives and acts. This makes learning and reasoning in this context inherently sequential (Howard, 1960). Active inference is a biologically plausible method operating on sequential problems[6] within a probabilistic framework, in which perception updates beliefs about hidden causes of observations, while action is selected to realize preferred and informative future observations (Friston et al., 2015, 2017; Parr et al., 2022).

Sequential Decision Processes

A standard basic formalism for sequential problem solving is the discrete-time *Markov decision process* (MDP; Bellman, 1957). An MDP can be represented as a tuple

$$\mathrm{MDP}\colon\ \langle \mathcal{S}, \mathcal{A}, T, R, \gamma \rangle\,, \tag{1.15}$$

where $\mathcal{S}$ and $\mathcal{A}$ denote discrete or continuous state and action spaces, respectively, therefore $\boldsymbol{s}(t) \in \mathcal{S}$, $\boldsymbol{a}(t) \in \mathcal{A}$. $T\left[\boldsymbol{s}(t+1) \mid \boldsymbol{s}(t), \boldsymbol{a}(t)\right]$ designates a transition distribution, $R\left[\boldsymbol{s}(t), \boldsymbol{a}(t)\right]$ a reward function, and $0 \leq \gamma \leq 1$ a discount factor. The *Markov property* mentioned extensively in Sections 1.3.1 and 1.3.2 as an attribute of the BN

[6] Another commonly used weakly bio-plausible framework for sequential problem solving is RL (Sutton & Barto, 2018); in this research, however, we borrow concepts from it only marginally.

relationships, applies naturally within this framework as well, i.e.,

$$p\left[\boldsymbol{s}(t+1) \mid \boldsymbol{s}(t), \boldsymbol{a}(t), \boldsymbol{s}(t-1), \boldsymbol{a}(t-1), \ldots\right] = p\left[\boldsymbol{s}(t+1) \mid \boldsymbol{s}(t), \boldsymbol{a}(t)\right]. \tag{1.16}$$

Within an MDP, the current state information is assumed to be completely and perfectly available to the agent.

However, in realistic use cases and environments, the full-observability assumption is too strong, as the agent typically receives only partial and noisy observations. A *partially-observable Markov decision process* (POMDP; Åström, 1965; Smallwood and Sondik, 1973) extends an MDP to those cases:

$$\text{POMDP}\colon \left\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega, R, \gamma \right\rangle, \tag{1.17}$$

where additionally $\mathcal{O}$ denotes an observation space, and $\Omega\left[\boldsymbol{o}(t) \mid \boldsymbol{s}(t), \boldsymbol{a}(t)\right]$ an observation distribution, with the observation $\boldsymbol{o}(t) \in \mathcal{O}$ mediating the full state information $\boldsymbol{s}(t) \in \mathcal{S}$ at the same time-point (Kaelbling et al., 1998; Spaan, 2012; Lauri et al., 2023). Since the true state $\boldsymbol{s}(t)$ is hidden, the agent maintains internal *belief state*, i.e., a posterior distribution over the hidden state $\boldsymbol{s}(t)$:

$$b_t(\boldsymbol{s}) \equiv p\left[\boldsymbol{s}(t) = \boldsymbol{s} \mid \boldsymbol{o}(1\colon t), \boldsymbol{a}(1\colon t-1)\right]. \tag{1.18}$$

For a more realistic continuous state space, the belief update can be recursively formulated as

$$b_{t+1}\left(\boldsymbol{s'}\right) \triangleq \eta\, \Omega\left[\boldsymbol{o}(t+1) \mid \boldsymbol{s'}, \boldsymbol{a}(t)\right] \int_{\mathcal{S}} T\left[\boldsymbol{s'} \mid \boldsymbol{s}, \boldsymbol{a}(t)\right] b_t(\boldsymbol{s}) d\boldsymbol{s}, \tag{1.19}$$

where η is a normalizing constant (Kaelbling et al., 1998). Thus, although the agent does not directly observe the actual state, the belief state summarizes the action-observation history sufficiently for sequential decision-making (Kaelbling et al., 1998).

Active Inference

Active inference is a framework for perception, learning, and action grounded in variational Bayesian inference and the free-energy principle (Friston et al., 2015, 2017; Parr et al., 2022). Instead of treating action selection primarily as reward maximization (as in, e.g., RL), active inference assumes that an agent possesses a *generative model* relating hidden states $\boldsymbol{s} \in \mathcal{S}$, actions $\boldsymbol{a} \in \mathcal{A}$ (or policies), and observations $\boldsymbol{o} \in \mathcal{O}$. Perception corresponds to inferring the hidden states that best explain sensory observations under this model, i.e., $p(\boldsymbol{s} \mid \boldsymbol{o}, \mathcal{M})$.

In PGMs in general, exact Bayesian inference is computationally difficult; specifically, in BNs, probabilistic inference is NP-hard (Cooper, 1990), as well as its approximation with guaranteed accuracy (Dagum & Luby, 1993). This is the case for many non-trivial real-world inference problems. Practical inferential systems, therefore, often rely on approximate inference methods, such as sampling-based or variational approaches, which compromise inferential exactness and theoretical guarantees for tractability in appropriate model classes (Blei et al., 2017).

The active inference framework works within the variational paradigm as, instead of computing computationally expensive exact posterior distribution over the hidden states $p(\boldsymbol{s} \mid \boldsymbol{o}, \mathcal{M})$, it introduces an approximate posterior $q(\boldsymbol{s} \mid \mathcal{M})$ optimized by minimizing *variational free energy* (VFE):

$$F[q, \boldsymbol{o}] \triangleq \mathbb{E}_{q(\boldsymbol{s})}\big[\log q(\boldsymbol{s}) - \log p(\boldsymbol{o}, \boldsymbol{s})\big], \tag{1.20}$$

where $q(\boldsymbol{s})$ is a notational shortcut for $q(\boldsymbol{s} \mid \mathcal{M})$ representing the model's belief, and $p(\boldsymbol{o}, \boldsymbol{s})$ corresponds to the generative model. By this definition, VFE upper-bounds the negative log evidence $-\log p(\boldsymbol{o})$, and minimizing it makes $q(\boldsymbol{s})$ approximate the true posterior $p(\boldsymbol{s} \mid \boldsymbol{o})$ (Friston, 2010; Friston et al., 2015; Parr et al., 2022). In this sense, perception in active inference is implemented as approximate Bayesian belief updating.

Similarly, action selection can be formulated as a free-energy minimization as well. In contrast with RL, the evaluation of a *policy*[7] $\pi \in \mathfrak{P}$ is not based on externally supplied rewards, but on its *expected free energy* (EFE) $G(\pi)$ expected to be minimized:

$$\pi^* \triangleq \operatorname*{argmin}_{\pi \in \mathfrak{P}} G(\pi), \tag{1.21}$$

where

$$G(\pi) \triangleq -\mathbb{E}_{q(\hat{\boldsymbol{o}}|\pi)}\big[D_{\mathrm{KL}}\left[q(\hat{\boldsymbol{s}} \mid \hat{\boldsymbol{o}}, \pi) \,\|\, q(\hat{\boldsymbol{s}} \mid \pi)\right]\big] \tag{1.22a}$$

$$-\mathbb{E}_{q(\hat{\boldsymbol{o}}|\pi)}\big[\log p(\hat{\boldsymbol{o}} \mid C)\big], \tag{1.22b}$$

with $\hat{\boldsymbol{s}}$ and $\hat{\boldsymbol{o}}$ denoting predicted future trajectories of hidden states and observations, respectively, $q(\hat{\boldsymbol{s}} \mid \pi)$ describing beliefs about future state trajectories predicted under policy π, $q(\hat{\boldsymbol{s}} \mid \hat{\boldsymbol{o}}, \pi)$ describing the beliefs that would result after obtaining future observations, and $p(\hat{\boldsymbol{o}} \mid C)$ encoding prior preferences over outcomes. The first term (Eq. 1.22a) then favors policies expected to reduce uncertainty about hidden states (i.e., producing epistemic value), while the second (Eq. 1.22b) favors policies expected to produce preferred observations (i.e., generating pragmatic value) (Parr et al., 2022, Section 2.8). Consequently, policy selection in active inference jointly accounts for epistemic and pragmatic value. This makes active inference related to BED (Section 1.3.2), since both score candidate actions or experiments in part by their expected epistemic value.

1.3.4 Neurosymbolic and Multilevel Architectures

The formalisms introduced in Sections 1.3.1 to 1.3.3 describe structured probabilistic, causal, and decision-theoretic reasoning. In embodied agents operating in realistic environments, however, such reasoning methods usually cannot operate directly on raw

[7]Note that within the active inference theory, we understand a policy as a sequence of actions (Parr et al., 2022, p. 31), as opposed to the RL paradigm where a policy is defined as a state-action mapping.

sensory inputs. Visual, proprioceptive, and motor data, for instance, are continuous, often noisy, and high-dimensional, while reasoning often requires more abstract entities such as objects, relations, events, actions, goals, and causal variables (Harnad, 1990; Lake et al., 2016; Gärdenfors, 2019; Butz et al., 2025; Schölkopf et al., 2021). Thus, a representational gap commonly exists between non-symbolic perception and higher-level reasoning in artificial systems. *Neurosymbolic* approaches address it by combining neural or otherwise subsymbolic learning and inference mechanisms with explicit symbolic, logical, graphical, probabilistic, or causal structures (Garcez et al., 2019; Besold et al., 2021).

Within neurosymbolic architectures, the subsymbolic component is typically responsible for learning from continuous data, constructing representations (Bengio et al., 2013; van den Oord et al., 2017), or estimating non-trivial functions (Hornik et al., 1989) and distributions (Papamakarios et al., 2021). The symbolic (or, more generally, structured) component, by contrast, provides explicit variables, relations, constraints, compositional structure, or causal mechanisms over which inference and planning can be performed. For example, differentiable probabilistic logic systems combine neural predictors with logical-probabilistic programs (Manhaeve et al., 2021), while causal representation learning aims to recover representations whose components support causal reasoning and transfer across environments (Schölkopf et al., 2021).

A related design principle is *multilevel representation*. In the cognitive and embodied context, rather than modeling an agent's environment at a single granularity, multilevel architectures organize information across multiple abstraction levels: for example, from low-level sensorimotor states or observations, through object- or event-level descriptions, to higher-level relational or causal structures. Lower levels preserve perceptual and metric detail, while the higher levels support more compact reasoning, generalization, and planning. The connections between levels can then be understood as an abstraction–refinement relation: lower-level states are mapped into higher-level descriptions, while higher-level decisions constrain or guide lower-level perception and action. This kind of hierarchical organization in knowledge representation and cognitive processing is biologically plausible, as in the case of the neural organization of the prefrontal cortex (Badre et al., 2010; Miller & Cohen, 2001) and of different sensory cortices (Rolls, 2016; Hubel & Wiesel, 1962), as well as in cognitive models in general (Broadbent, 1977; Newell, 1994; Cohen, 2000; Balleine et al., 2015; Eckstein & Collins, 2021). A similar principle has been adopted in robotics in refinement-based architectures such as REBA (Sridharan et al., 2019) and REBA-KRL (Sridharan, 2020), in knowledge-driven architectures that integrate symbolic reasoning with learning and probabilistic decision-making (Mota et al., 2021; Dodampegama & Sridharan, 2023), and in task-and-motion planning systems, which similarly coordinate symbolic task planning with geometric or motion-level feasibility reasoning (Cambon et al., 2009; Kaelbling & Lozano-Pérez, 2011; Srivastava et al., 2014; Garrett et al., 2021).

Chapter 2

Related Work

The preceding chapter introduced the theoretical concepts used in this thesis. This chapter situates the proposed architecture within two major groups of related computational work: sensorimotor world models (Section 2.1) and computational visual perception (Section 2.2). The goal is not to survey these areas exhaustively, but to identify the approaches most closely related to the proposed combination of object-centric perception, sensorimotor action-effect learning, and structured neurosymbolic modeling.

2.1 Sensorimotor World Models

A first relevant line of work comes from developmental robotics and sensorimotor representation learning, where embodied agents acquire regularities through self-generated action (Cangelosi & Schlesinger, 2015; Asada et al., 2009). Motor babbling provides an early mechanism for collecting action-sensory data (Saegusa et al., 2009), while goal babbling biases exploration toward reachable outcomes (Rolf et al., 2010; Baranes & Oudeyer, 2013). These strategies are often combined with intrinsically motivated exploration, in which internal learning signals guide the agent toward novelty, competence progress, or information gain (Oudeyer & Kaplan, 2007). Such systems are closely aligned with the developmental motivation of this thesis, since they treat action as a source of learning rather than only as an output of a controller. Surveys of body representation learning emphasize related mechanisms for acquiring an active self, body schema, and sensorimotor prediction (Schillaci et al., 2016; Nguyen et al., 2021; Gama, 2026). The present thesis follows this developmental orientation but differs by emphasizing an explicitly structured world model (WM): the learned action-effect relation is represented as a graph of variables and local mechanisms rather than as an implicit forward or inverse mapping alone.

A second related line is model-based reinforcement learning with neural world models. The world-model architecture of Ha and Schmidhuber (2018) demonstrated that compact latent-dynamics models can support policy learning within an internally sim-

ulated environment. PlaNet learned latent dynamics directly from pixel observations and used the learned model for online planning (Hafner et al., 2019). Later Dreamer variants trained policies on imagined trajectories in a learned latent space, spanning diverse visual and proprioceptive control domains (Hafner et al., 2025). These systems show the practical strength of learned predictive models for control. However, their main objective is usually task performance under a reward or policy-learning objective. Their latent states are optimized for control and prediction, but are not necessarily organized as inspectable causal variables, object records, or grounded conceptual domains. By contrast, the architecture proposed here prioritizes interpretability, causal structure, and compatibility with perceptual representations, even at the cost of using a much smaller proof-of-concept implementation.

Recent work has also begun to connect world models with object-centric representations. In object-centric robotic manipulation, embodied perceptual learning, policy learning, and task-oriented object interaction are commonly treated as coupled components of manipulation systems (Zheng et al., 2025). More directly, FOCUS combines world-model-based reinforcement learning with object-level representations for robotic manipulation (Ferraro et al., 2025). This is close to the present thesis in its treatment of objects as important units for predictive control. The distinction is again primarily architectural and explanatory: the present work does not attempt to learn a high-performing manipulation policy but instead articulates how perceptual object descriptors, proprioceptive variables, causal mechanisms, and action inference could fit into a single neurosymbolic perception-action architecture.

A third group of related work concerns causal and active world modeling. Causal representation learning argues that robust generalization requires representations whose components correspond to causally meaningful variables and mechanisms (Schölkopf et al., 2021). BISCUIT illustrates this direction by learning causal representations from binary interactions, where interaction changes provide information about the underlying causal factors (Lippe et al., 2023). In robotics, causal modeling is relevant because actions can be interpreted as interventions on the agent-environment system rather than as passive observations (Hellström, 2021). Active Bayesian causal inference (ABCI) similarly treats actions or experiments as information-seeking interventions for learning causal structure or mechanisms (Toth et al., 2022). The WM proposed in this thesis adopts this motivation by representing actions as intervention-like assignments in a structured graph. Nevertheless, the implemented experiment remains deliberately weaker than full causal discovery: the graph is manually specified, the mechanisms are simple, and the evaluation is limited to offline proprioceptive transitions.

2.2 Computational Visual Perception

The perceptual part of this thesis is related first to active and interaction-driven computer vision. Active vision treats perception as a process that can be improved by controlling viewpoint, attention, or interaction rather than passively processing fixed images (Bajcsy, 1988; Aloimonos et al., 1988). In robotics, early work showed that object boundaries can be discovered through manipulation-induced motion (Fitzpatrick et al., 2003; Fitzpatrick, 2003). Later interaction-based approaches learned object properties through manipulation (Lohmann et al., 2020; Kruzliak et al., 2024), discovered articulated object parts through action (Gadre et al., 2021), or modeled object kinematics from interaction (Nematollahi et al., 2020). This interaction-centered view is closely related to the architecture proposed here. However, the implemented perceptual experiments in this thesis do not yet use the agent's own actions to actively reveal object structure. Instead, they evaluate whether recorded RGB-D observations can be transformed into object-centric descriptors that could later be used in such a closed perception-action setting.

Modern computer vision provides powerful components for object detection, segmentation, tracking, correspondence estimation, and pose-related processing. Segment Anything and its video extension provide general object-level image and video supports (Kirillov et al., 2023; Ravi et al., 2024), while YOLO-style open-vocabulary detection and segmentation can provide practical real-time object hypotheses (Wang et al., 2025). Tracking-by-detection methods such as SORT and BoT-SORT support short-term identity maintenance through detection association (Bewley et al., 2016; Aharon et al., 2022). Recent local-feature and matching approaches, such as DeDoDe and LightGlue, support geometric correspondence estimation under difficult visual conditions (Edstedt, Bökman, Wadenbäck, & Felsberg, 2024; Lindenberger et al., 2023). The perceptual implementation in this thesis uses such methods instrumentally. Its contribution is not a new segmentation, tracking, or pose-estimation algorithm, but the organization of their outputs into object records, retained evidence, diagnostic pose information, conceptual projections, and scene-level descriptors.

A closely related research area is object-centric representation learning. Slot Attention decomposes visual input into object-like latent slots (Locatello et al., 2020). Extensions to video, such as SAVi++, learn object-centric representations of dynamic scenes from richer visual prediction signals (Elsayed et al., 2022). Newer concept-binding approaches attempt to connect neural object-centric representations with more interpretable conceptual structure (Stammer et al., 2024). These approaches are relevant because they address the same representational pressure: downstream reasoning benefits from representations organized around objects rather than undifferentiated image-level embeddings. The present thesis, however, employs a more transparent, handcrafted perceptual representation. Visual and kinematic descriptors are not ex-

pected to match the invariance or scalability of learned object-centric models, but they make the intermediate representational structure inspectable and easier to connect to conceptual-space and WM variables.

The proposed perceptual subsystem is also related to qualitative spatial reasoning (QSR) and scene-graph approaches. QSR formalisms explicitly represent spatial and topological relations, for example, through region-based or point-set relations (Egenhofer & Franzosa, 1991; Randell et al., 1992; Cohn & Renz, 2008). Scene-graph approaches similarly represent images as objects and explicit semantic or spatial relations, which support retrieval, reasoning, and structured visual understanding (Johnson et al., 2015; Krishna et al., 2017; Chang et al., 2023). Such relational representations are important for connecting perception to reasoning and action, but the current implementation includes only a structured scene descriptor and not a full relational scene parser. In this sense, the visual part of the thesis occupies an intermediate position: it is more structured and object-centric than a monolithic visual embedding, but less expressive than a complete scene graph or a qualitative spatial model. This limitation is intentional within the proof-of-concept scope and motivates the later integration of perceptual variables into the structured WM.

Chapter 3

Aims and Problem Formulation

In this chapter, we formulate the specific aims of our work, along with the simplifying assumptions and delimitations applied (Section 3.1), formulate the desired attributes for the proposed system based on the theoretical concepts highlighted in Chapters 1 and 2 (Section 3.2), and define the problem formalization and additional related nomenclature used throughout the thesis (Section 3.3).

3.1 Scope

As stated in Introduction, the present thesis proposes a minimal neurosymbolic cognitive architecture that enables an embodied agent to autonomously learn and reason about its environment. More specifically, we aim to provide a theoretical design and partial implementation of a perception-action architecture that connects grounded perceptual representations, causal world modeling, and active sensorimotor exploration.

To elaborate on Introduction, this work follows three primary objectives:

1. We design and develop a perceptual system that processes dynamic visual input into object-centric conceptual representations and scene descriptors (Section 4.2).
2. We design and develop a complementary sensorimotor world model (WM) that learns structured action-effect regularities from embodied interaction (Section 4.3).
3. We formulate both components within a common perception-action setting, so that perception, inference, and action generation can be treated as coupled parts of a single embodied PAL (Section 4.1).

These objectives follow from the developmental and formal framing introduced in Chapter 1. Specifically, sensorimotor cognition motivates treating action as part of the agent's learning rather than only as an output of a controller (Section 1.1); visual perception motivates an object-centric representational layer between raw sensory input and higher-level reasoning (Section 1.2); and probabilistic, causal, and active-inference formalisms motivate an explicit WM that can support uncertainty-sensitive

prediction, inference, and intervention-like action selection (Section 1.3). The proposed architecture is therefore scoped to the minimal components needed to model a PAL operating a robotic agent under unstructured, noisy, and high-dimensional data, similar to real-world settings. More detailed architecture requirements and their justification are listed in Section 3.2.

Following the theoretical design (Chapter 4), selected essential components of the architecture are implemented and evaluated to test their feasibility. Note that the scope of this work is intentionally restricted to a proof-of-concept implementation demonstrating the basic functionality of these components. The thesis does not, however, intend to produce a general autonomous robotic framework, solve unrestricted open-world perception, or fully model the selected cognitive functions from which the architecture is inspired. Instead, we examine here whether selected cognitive and computational principles can be integrated into an inspectable and interpretable system that supports basic perception, causal learning, state inference, and action generation under controlled conditions.

Accordingly, higher-level capacities suggested as potential corollaries of the presented architecture, such as language, social cognition, and abstract commonsense reasoning, are outside the scope of the implementation and experimental validation. They should be considered solely as possible extensions of the proposed framework, and will be discussed later in Section 6.4. The formal assumptions about the environment, observations, actions, and modeled variables are specified separately in Section 3.3.

3.2 Architectural Desiderata

Given the scope defined above, the proposed architecture should satisfy several desiderata derived from the theoretical background in Chapter 1. Here, we do not yet specify the concrete implementation details; instead, we use the following requirements to constrain the architecture developed in Chapter 4.

Embodied Interaction As discussed in Section 1.1, sensorimotor cognition is grounded in the coupling between action, bodily state, and sensory consequences (Varela et al., 1991; O'Regan & Noë, 2001; Noë, 2004). Hence, the agent should not be modeled as a passive observer consuming independent observations, but as an embodied system whose actions change both its own sensorimotor state and the observations available for learning.

Grounded Perceptual Representation Raw visual input is a high-dimensional sensory signal from which object identity, spatial relations, and action-relevant structure are not directly given; visual perception is therefore commonly treated as an inverse or inferential problem under ambiguity and noise (Marr, 2010; Kersten et al., 2004; Yuille & Kersten, 2006). Following Section 1.2, the architecture should include

a perceptual representation layer that transforms visual input into structured object-centric descriptions preserving relevant visual, spatial, and kinematic regularities (Harnad, 1990; Barsalou, 1999; Vogt, 2002; Gärdenfors, 2000, 2014). These representations should be usable by downstream learning and reasoning mechanisms.

Uncertainty Representation Since the agent operates under partial observability and receives incomplete and noisy observations, its internal states, object properties, and action consequences cannot generally be treated as perfectly known. The WM should explicitly account for uncertainty and should support probabilistic inference over unobserved or uncertain variables, following the probabilistic modeling principles introduced in Section 1.3.1 (Pearl, 1988; Koller & Friedman, 2009).

Causal World Modeling Probabilistic dependence alone is insufficient for modeling how an embodied agent changes its environment through action. Following the causal formalism introduced in Section 1.3.2, the WM should represent structured action-effect relations in a way that distinguishes, where possible, causal mechanisms from statistical associations (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021). Thus, the agent's actions are treated as intervention-like manipulations constrained by the agent's embodiment.

Active Exploration and Action Generation The agent should be able to select actions not only to reach externally specified goals, but also to obtain informative observations and reduce uncertainty about its own model. This follows from the links between autonomous experimental design, active inference, and epistemic value discussed in Sections 1.3.2 and 1.3.3 (Friston et al., 2015, 2017; Parr et al., 2022; Toth et al., 2022). Consequently, action generation should be formulated as part of the inferential loop rather than as an independent controller attached after learning.

Neurosymbolic Structure The architecture must bridge continuous sensorimotor data, perceptual abstractions, and structured variables usable for reasoning. Following Section 1.3.4, this system attribute requires a neurosymbolic and multilevel organization that combines subsymbolic learning components with explicit structured representations (Garcez et al., 2019; Besold et al., 2021; Lake, 2014; Lake et al., 2016; Butz et al., 2025). The resulting system should remain modular, inspectable, and interpretable, since the present thesis aims not only at task performance, but also at an explicit computational account of how grounded perception, causal world modeling, and action can be integrated.

3.3 Problem Formalization

In accordance with Sections 3.1 and 3.2, we formalize the agent–environment interaction as a discrete-time, finite-horizon, partially observable controlled process (Spaan, 2012). Since the agent is expected to operate in a realistic, stochastic, and not fully observable environment, with inference performed under uncertainty, we model the setting after a POMDP[8] (Åström, 1965; Smallwood & Sondik, 1973; Kaelbling et al., 1998; Lauri et al., 2023). In the context of this work, however, we do not adopt the full standard POMDP formalism formulated as an RL problem, since several of its assumptions are too restrictive for our use case and we do not primarily operate within the RL paradigm. In particular, we do not define a task-specific reward function, discount factor, or policy-learning objective at this level. Instead, we use a reduced POMDP-like formalism to define the state, action, observation, and transition structure within which the proposed architecture operates.

Let us define the environment

$$\mathcal{E} \triangleq \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, \Omega, h \rangle \,, \tag{3.1}$$

where $h \in \mathbb{N}^+$ denotes a finite time horizon, with an environmental discrete time step $0 \leq t \leq h$, and the environment starting at time $t = 0$. The state space is defined as $\mathcal{S} \triangleq \mathbb{R}^{d_s}$, with d_s being the number of state attributes or variables, and $\boldsymbol{s}(t) \in \mathcal{S}$ denoting the state vector at time t. The action space is defined as $\mathcal{A} \triangleq \mathbb{R}^{d_q}$, with d_q denoting the number of the agent's controllable degrees of freedom (DoF) or actuators, and $\boldsymbol{a}(t) \in \mathcal{A}$ being an action vector executed at time t.

The observation space is decomposed into visual and proprioceptive components:

$$\mathcal{O} \triangleq \mathcal{O}^{\text{vis}} \times \mathcal{Q}, \tag{3.2}$$

where

$$\mathcal{O}^{\text{vis}} \triangleq [0, 1]^{M \times N \times 4 \times c} \tag{3.3}$$

denotes the visual observation space, and $\mathcal{Q} \triangleq \mathbb{R}^{d_q}$ denotes the proprioceptive space of the agent's joint configurations. Then, an observation at time t is written as

$$\boldsymbol{o}(t) \triangleq [\boldsymbol{v}(t), \boldsymbol{q}(t)] \in \mathcal{O}, \tag{3.4}$$

where $\boldsymbol{v}(t) \in \mathcal{O}^{\text{vis}}$ is a visual percept formalized as a sequence of $c \in \mathbb{N}^+$ unit-normalized RGB-D frames capturing the agent's recent visual experience, and $\boldsymbol{q}(t) \in \mathcal{Q}$ denotes the agent's joint configuration. At the beginning of an episode, when fewer than c observational frames are available, the visual history may be truncated or padded.

The transition model is given by

$$T \left[\boldsymbol{s}(t+1) \mid \boldsymbol{s}(t), \boldsymbol{a}(t) \right] \tag{3.5}$$

[8] For a review of standard POMDP formalism and the related concepts, see Section 1.3.3

denoting the stochastic state transition distribution over the next state after executing $\boldsymbol{a}(t)$ in $\boldsymbol{s}(t)$. Similarly, the observation model is given by

$$\Omega\left[\boldsymbol{o}(t) \mid \boldsymbol{s}(t), \boldsymbol{a}(t)\right]. \tag{3.6}$$

The time indexing follows Section 1.3.3; thus, $\boldsymbol{o}(t)$ denotes the observation associated with the interaction step at state $\boldsymbol{s}(t)$ under action $\boldsymbol{a}(t)$. Since $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ are continuous in this formulation, T and Ω are understood as stochastic kernels, or as probability densities where appropriate. In the implemented architecture, however, neither T nor Ω is specified explicitly over the full $\mathcal{S}$ and $\mathcal{O}$, as such models would be intractably complex in the considered setting. Instead, selected components of these functions are approximated by the perceptual subsystem and the learned sensorimotor WM.

We assume that the agent's true joint configuration is available at each time step. Thus, for the sake of simplicity, $\boldsymbol{q}(t)$ is both observed and treated as a component of the underlying state, i.e., $\boldsymbol{q}(t) \subset \boldsymbol{s}(t)$ as well as $\boldsymbol{q}(t) \subset \boldsymbol{o}(t)$, with $\mathcal{Q} \subset \mathcal{S}$ and $\mathcal{Q}$ forming the proprioceptive factor of $\mathcal{O}$. By contrast, the visual percept $\boldsymbol{v}(t)$ is not itself a state vector, but a raw measurement of the environment. The perceptual subsystem is expected to process $\boldsymbol{v}(t)$ into estimated structured state descriptors $\hat{\boldsymbol{s}}^{\text{vis}}(t)$ approximating only a subvector of the true state, i.e.,

$$\hat{\boldsymbol{s}}^{\text{vis}}(t) \approx \boldsymbol{s}^{\text{vis}}(t) \subset \boldsymbol{s}(t). \tag{3.7}$$

Hence, $\mathcal{O} \not\subseteq \mathcal{S}$.

For the purposes of this thesis, actions are simplified as commanded joint displacements for $0 \leq t < h$:

$$\boldsymbol{a}(t) \triangleq \boldsymbol{q}(t+1) - \boldsymbol{q}(t). \tag{3.8}$$

This corresponds to simple positional actuator control and abstracts away from more complex motor dynamics, biomechanics, and low-level control.

Compared to a standard finite-horizon POMDP, commonly represented as a 9-tuple (Spaan, 2012; Lauri et al., 2023)

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, T_0, \Omega, R, \gamma, h \rangle, \tag{3.9}$$

our formulation intentionally omits an initial state distribution T_0, a reward function R, and a discount factor γ. T_0 is not central to the architectural problem considered here, while R and γ-discounting are tied to specific externally defined tasks. Since the thesis focuses on continual embodied learning, exploration, and model construction rather than on optimizing a fixed task objective, these terms are not included in the environment definition. Task-specific preferences or exploration objectives are introduced later at the level of the agent and its WM, rather than as part of the basic environment formalization.

Chapter 4

Proposed Architecture

In this chapter, we specify the architecture proposed to address the aims and requirements outlined in Chapter 3. The chapter forms the main theoretical contribution of this thesis. We first describe the overall system organization (Section 4.1), then detail the visual-perceptual subsystem (Section 4.2), and finally specify the neurosymbolic world model (Section 4.3). The architectural design is presented here as a general formulation. Some components were implemented and evaluated in the proof-of-concept experiments reported in Chapter 5, whereas others are included as architectural requirements or proposed extensions. Unless a component is explicitly marked as implemented and evaluated, it should be understood as a design-level specification rather than an empirically validated claim.

4.1 Overall System Architecture

At the top level, the proposed architecture operates within the formal setting defined in Section 3.3 as a structured, modular PAL (Fig. 4.1). The agent (i.e., the embodiment of the architecture) receives observational input from the environment in two streams: visual observations are processed by the *perceptual subsystem*, which maps high-dimensional visual input to structured object-centric scene descriptors while maintaining internal conceptual representations aiding perception; proprioceptive observations provide information about the agent's bodily configuration and are passed directly to the *world-model subsystem* (WM). Together, these inputs form the structured approximate state representation used by the WM to learn and represent dependencies between actions, state descriptors, and sensory consequences, and to generate further actions that close the PAL.

Note that this division is functional rather than purely implementational: the subsystems are specified by their roles in the PAL, while concrete implementation choices may instantiate these roles through different module boundaries and data structures.

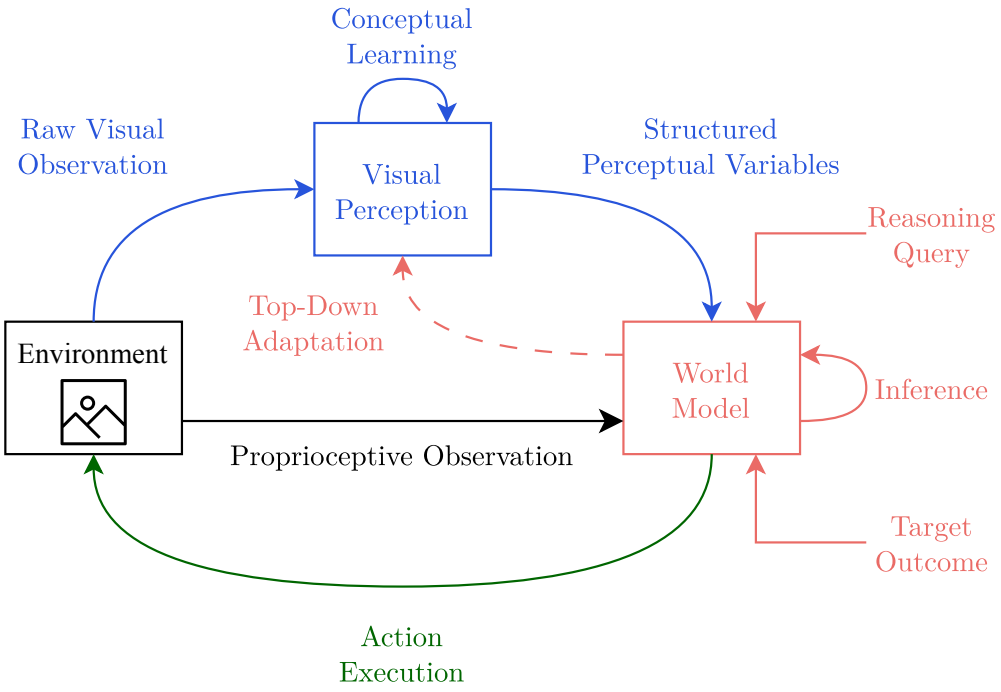


Figure 4.1: High-level organization of the proposed perception-action architecture. Visual observations are transformed into structured perceptual representations, combined with proprioceptive state information, and integrated into a neurosymbolic WM supporting inference and action generation within a closed embodied interaction loop (cf. Fig. 1.1).

Formally, the architecture operates on observations (Eq. 3.4)

$$\boldsymbol{o}(t) \triangleq [\boldsymbol{v}(t), \boldsymbol{q}(t)] \in \mathcal{O}, \tag{4.1}$$

where $\boldsymbol{v}(t) \in \mathcal{O}^{\text{vis}}$ is the raw visual observation and $\boldsymbol{q}(t) \in \mathcal{Q}$ is the directly observed proprioceptive component. The visual stream is processed by the perceptual subsystem, which estimates a visually accessible subvector of the environmental state:

$$\Pi_{\boldsymbol{\theta}^{\text{perc}}}[\boldsymbol{v}(t)] = \hat{\boldsymbol{s}}^{\text{vis}}(t) \approx \boldsymbol{s}^{\text{vis}}(t) \subset \boldsymbol{s}(t). \tag{4.2}$$

Here, $\Pi_{\boldsymbol{\theta}^{\text{perc}}}$ denotes the parametrized perceptual process, and the parametrization $\boldsymbol{\theta}^{\text{perc}}$ may capture both learned perceptual parameters and possible top-down modulation from the rest of the architecture (McMains & Kastner, 2011). For details of the perceptual subsystem, see Section 4.2.

The proprioceptive component is not inferred from the visual stream. Instead, as per Section 3.3, it is assumed to be directly available to the agent and treated both as part of the observation and as a component of the underlying state, i.e., $\boldsymbol{q}(t) \subset \boldsymbol{o}(t)$ and $\boldsymbol{q}(t) \subset \boldsymbol{s}(t)$. The *structured state estimate* available to the WM is therefore given by

$$\hat{\boldsymbol{s}}(t) = \hat{\boldsymbol{s}}^{\text{vis}}(t) \cup \boldsymbol{q}(t), \tag{4.3}$$

where $\cup$ denotes structured composition or vector concatenation. In general, $\hat{\boldsymbol{s}}(t)$ remains a partial and approximate state estimate, since it contains directly observed proprioceptive variables and visually inferred descriptors, but not necessarily all variables constituting the underlying environmental state.

The WM operates on the structured *sensorimotor history*[9]

$$\mathcal{H}_t \triangleq \big[\hat{\boldsymbol{s}}(0), \boldsymbol{a}(0), \hat{\boldsymbol{s}}(1), \boldsymbol{a}(1), \ldots, \boldsymbol{a}(t-1), \hat{\boldsymbol{s}}(t)\big], \tag{4.4}$$

which serves as evidence for updating an internal model $\mathcal{M}^{\mathrm{wm}}(t)$ parametrized by $\boldsymbol{\theta}^{\mathrm{wm}}$. At this level of description, $\mathcal{M}^{\mathrm{wm}}(t)$ is treated as an abstract learned model of structured dependencies between actions, state descriptors, and sensory consequences, rather than as a full transition model over $\mathcal{S}$.

The model supports prediction, state inference, and action generation conditioned on the available state estimate $\hat{\boldsymbol{s}}(t)$, the current model state, and target values provided externally or generated internally through exploration. The resulting action $\boldsymbol{a}(t) \in \mathcal{A}$ is executed in the environment, changing the agent–environment state and producing a new observation at the next interaction step, thereby closing the PAL. For a more detailed formalization and description of the WM, see Section 4.3.

4.2 Visual Perception and Semantic Representations

The perceptual subsystem (Fig. 4.2) instantiates the visual component of the architecture by transforming raw RGB-D observations into structured *object-centric descriptors* and *conceptual representations*. Its role, as mentioned in Section 4.1, is not to reconstruct the complete environmental state, but to estimate visually accessible state variables, maintain object-level perceptual memory, and organize visual properties and motion patterns into representations usable by the WM.

This section first formalizes in more detail the perceptual mapping introduced in Section 4.1 (Section 4.2.1), then describes the processing stages through which visual input is discretized, maintained, quantified, and composed into scene-level descriptions (Section 4.2.2). Finally, it relates these design choices to selected neurocognitive principles of active, object-centered, and concept-mediated perception (Section 4.2.3).

[9]This notion is closely related to robotic trajectories, such as the formulation used by Cibula et al. (2025). Here, however, the term *history* is used more broadly, since the ordered sequence may contain structured observations or inferred descriptors rather than only physical states and controls.

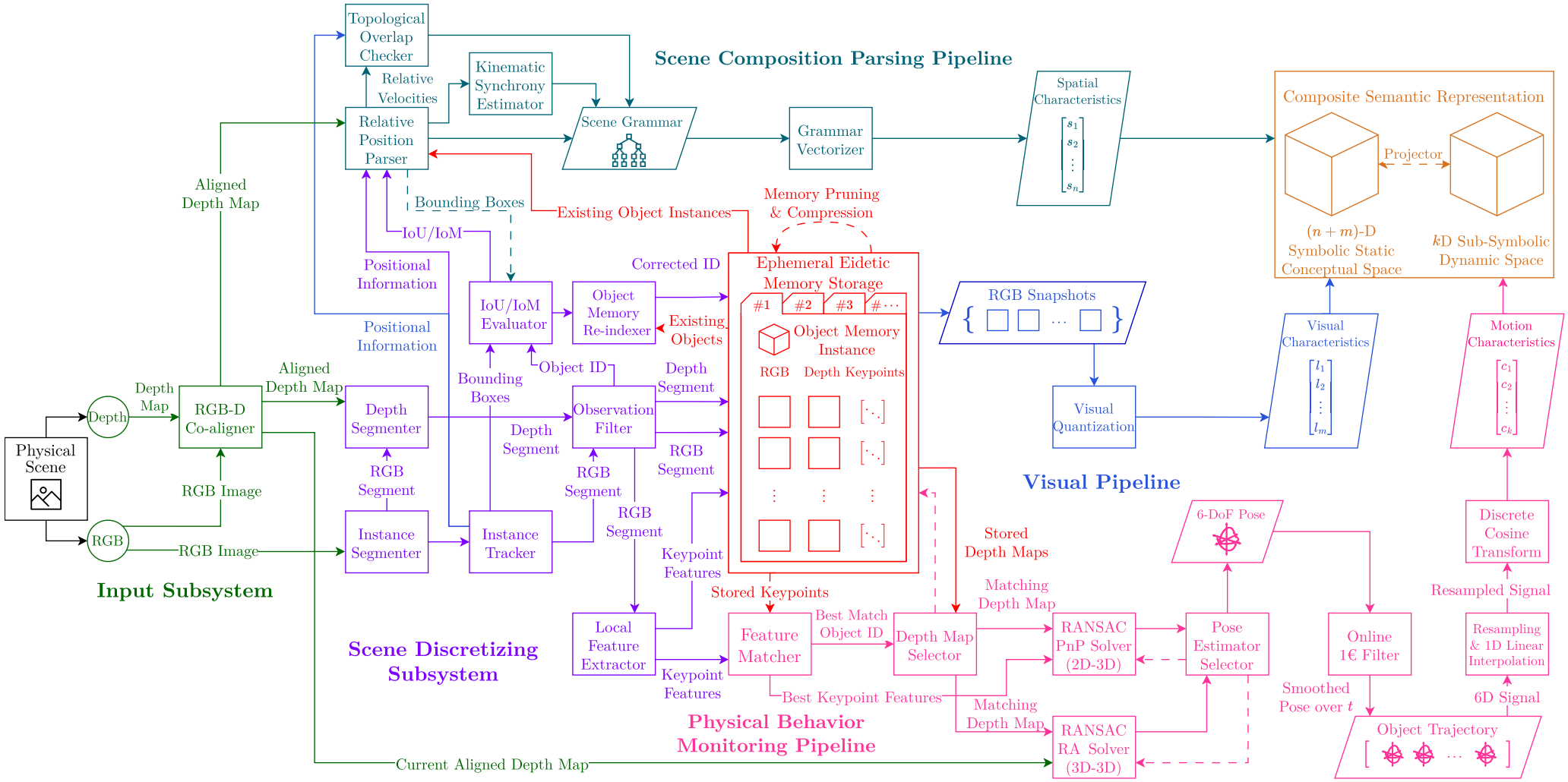


Figure 4.2: Full high-level diagram of the proposed perceptual architecture. The rectangular blocks correspond to either processing units or general elements; parallelograms denote data structures. Subsystem organization denoted by the colorization corresponds with the subsystem specifications in Section 4.2.2. Solid arrowed lines denote the forward data flow, while the dashed ones mark the feedback connections between the modules. As disclaimed in Sections 4.2 and 4.2.2, the full extent of the perceptual architecture illustrated in this figure, is primarily theoretical.

4.2.1 Formalization

As introduced in Section 4.1, the perceptual subsystem is formalized as a $\boldsymbol{\theta}^{\text{perc}}$-parametrized mapping from raw visual observation $\boldsymbol{v}(t) \in \mathcal{O}^{\text{vis}}$ to estimated visually accessible state descriptors $\hat{\boldsymbol{s}}^{\text{vis}}(t)$:

$$\Pi_{\boldsymbol{\theta}^{\text{perc}}} : \boldsymbol{v}(t) \mapsto \hat{\boldsymbol{s}}^{\text{vis}}(t). \tag{4.5}$$

As per Eq. 4.2, $\hat{\boldsymbol{s}}^{\text{vis}}(t)$ is partial and approximate w.r.t. $\boldsymbol{s}(t) \in \mathcal{S}$. In restricted experimental settings, however, the combination of visual descriptors $\hat{\boldsymbol{s}}^{\text{vis}}(t)$ and proprioception $\boldsymbol{q}(t)$ may cover all task-relevant modeled variables, but this should be understood as completeness relative to the chosen task and representation, not as full access to the environmental state $\boldsymbol{s}(t)$.

The perceptual function Π should not be identified with the POMDP observation model Ω (Eq. 3.6). While Ω describes the generative relation between states, actions, and observations, Π denotes an inferential process that maps strictly visual observations to structured descriptors usable by downstream modeling. Thus, Π is a component of the agent architecture, not a generative model of the environment $\mathcal{E}$.

The parametrization $\boldsymbol{\theta}^{\text{perc}}$ should be understood broadly as the current configuration of the perceptual process. It may include learned parameters, hyperparameters, or maintained state of perceptual modules, longer-term changes in conceptual representations, and short-term top-down modulation from the rest of the architecture. In this sense, $\Pi_{\boldsymbol{\theta}^{\text{perc}}}$ allows perception to be bidirectional (Dijkstra et al., 2017; McMains & Kastner, 2011): bottom-up visual evidence constrains the produced descriptors, while higher-level expectations, task context, or attentional settings may influence what is selected, maintained, or disambiguated (Deutsch & Deutsch, 1963; Neisser, 1967; Cowan, 1998; Démuth, 2013). The present implementation, described in Section 5.1, instantiates only selected aspects of this adaptation, but the architectural formulation keeps $\boldsymbol{\theta}^{\text{perc}}$ as the formal entry point through which perceptual learning and top-down influence can affect the visual processing.

The output of Π is object-centric. Schematically, the estimated visual state can be decomposed as

$$\hat{\boldsymbol{s}}^{\text{vis}}(t) \triangleq \hat{\boldsymbol{s}}^{\text{scene}}(t) \cup \left[\hat{\boldsymbol{s}}_i^{\text{obj}}(t)\right]_{i=1}^{n_t}, \tag{4.6}$$

where n_t denotes the number of represented objects, $\hat{\boldsymbol{s}}^{\text{scene}}(t)$ denotes scene-level descriptors, and $\hat{\boldsymbol{s}}_i^{\text{obj}}(t)$ denotes the structured descriptor of object i. These object descriptors may include visual, spatial, kinematic, and conceptual components, for example

$$\hat{\boldsymbol{s}}_i^{\text{obj}}(t) = \hat{\boldsymbol{s}}_i^{\text{app}}(t) \cup \hat{\boldsymbol{s}}_i^{\text{spat}}(t) \cup \hat{\boldsymbol{s}}_i^{\text{kin}}(t) \cup \boldsymbol{c}_i(t), \tag{4.7}$$

where $\boldsymbol{c}_i(t)$ denotes the conceptual representation of object i in the relevant subspace of the perceptual subsystem's conceptual space. Note that this decomposition is abstract: the exact descriptor fields depend on the implemented perceptual modules, but the

architectural requirement is that the output remains structured, object-indexed, and usable by the WM.

4.2.2 Subsystems

The internal structure of Π formalized above can be seen in Fig. 4.2, which specifies the full extent of the perceptual architecture, including components implemented in the present proof of concept (Section 5.1) as well as those which have been purely theoretically designed. Overall, the perceptual system maps an aligned RGB-D stream to object observations, object-level visual and kinematic descriptors, spatial scene descriptors, and conceptual-space representations.

The first architectural requirement is *object-centric discretization* (Section 4.2.2.1). As downstream processing is expected to operate on structured variables rather than raw continuous sensory input, the perceptual system must identify persistent entities whose attributes, relations, and changes can be represented over time. After object discretization, the system separates perceptual representation into *appearance* (Section 4.2.2.4), *spatial* (Section 4.2.2.5), and *kinematic* (Section 4.2.2.3) branches. This organization is inspired by the need to dinstinguish what an object looks like, where it is situated relative to the agent and other objects, and how it behaves over time or under physical intervention. The capturing of these features is motivated by the distinction between visual processing for recognition and action (Goodale & Milner, 1992), and by affordance-oriented view that action-relevant object properties are revealed through possible interactions (Gibson, 2014).[10] Architecturally, the feature separation also prevents all perceptual information from being collapsed into a single opaque embedding.

4.2.2.1 Input Discretization

The input subsystem receives an RGB-D observation $\boldsymbol{v}(t) = [I(t), D(t)]$, where $I(t)$ and $D(t)$ are the RGB color image and the geometrically aligned depth map, respectively. A co-alignment step ensures that color and depth values are expressed in a shared image coordinate frame, so that object masks extracted from $I(t)$ can be used to index corresponding depth values in $D(t)$. The discretization stage can then be formalized as a *segmentation-tracking operator* Σ producing a set of object-level observations

$$\tilde{\mathcal{Z}}(t) = \Sigma_{\boldsymbol{\theta}^{\text{seg}}}\left[I(t)\right] = \left\{ \left[\tilde{\imath}_i(t), \tilde{\boldsymbol{b}}_i(t), \tilde{\boldsymbol{m}}_i(t) \right] \right\}_{i=1}^{\tilde{n}_t}, \tag{4.8}$$

where $\tilde{\mathcal{Z}}(t)$ denotes the set of candidate object instances at time t, $\boldsymbol{\theta}^{\text{seg}}$ denotes the parametrization of the segmenter-tracker, $\tilde{\imath}_i(t)$ is a candidate track identifier, $\tilde{\boldsymbol{b}}_i(t)$ is a bounding box, $\tilde{\boldsymbol{m}}_i(t)$ denotes an instance mask, and $\tilde{n}_t$ is the number of detected candidate instances.

[10] For more details on the cognitive basis of the human visual processing, see Section 1.2.1.

Since tracker identifiers and segmentation masks may fluctuate across frames, the candidate set $\tilde{\mathcal{Z}}(t)$ is further stabilized against the current *object memory* (see Section 4.2.2.2). Let $A_i(t)$ denote the image support of candidate object instance $1 \leq i \leq \tilde{n}_t$, given by its mask or bounding box, and let $A_j^{\mathrm{mem}}(t-1)$ denote the stored support of memory reference $1 \leq j \leq n_{t-1}^{\mathrm{mem}}$ from the previously maintained object memory. Candidate–memory compatibility is then evaluated using the standard intersection-over-union (IoU) overlap criterion (Everingham et al., 2010; Bewley et al., 2016; Aharon et al., 2022),

$$\mathrm{IoU}\left(A_i, A_j^{\mathrm{mem}}\right) \triangleq \frac{\left|A_i \cap A_j^{\mathrm{mem}}\right|}{\left|A_i \cup A_j^{\mathrm{mem}}\right|} \tag{4.9}$$

together with an intersection-over-minimum (IoM) auxiliary criterion,

$$\mathrm{IoM}\left(A_i, A_j^{\mathrm{mem}}\right) \triangleq \frac{\left|A_i \cap A_j^{\mathrm{mem}}\right|}{\min\left(\left|A_i\right|, \left|A_j^{\mathrm{mem}}\right|\right)}. \tag{4.10}$$

While IoU penalizes differences in support size and is therefore suitable for conservative association of similarly sized regions, IoM remains high when one region is largely contained in another. It is therefore useful for detecting partial masks, unstable segmentations, and containment-like overlaps that would receive a low IoU despite referring to the same physical object. Conversely, because IoM can overestimate compatibility when a small region lies inside a larger one, it is used as an auxiliary criterion rather than as a replacement for IoU.

The two criteria are combined by accepting a candidate–memory association whenever either $\mathrm{IoU}\left(A_i, A_j^{\mathrm{mem}}\right) \geq \tau_{\mathrm{IoU}}$ or $\mathrm{IoM}\left(A_i, A_j^{\mathrm{mem}}\right) \geq \tau_{\mathrm{IoM}}$, where τ_{IoU} and τ_{IoM} are association thresholds. This association rule supports short-term reference recovery and rejects inconsistent assignments, thereby reducing *reference hijacking*, where a newly detected or incorrectly tracked instance takes over the persistent memory reference of another object.

The resulting stabilized object-level observations are denoted by

$$\mathcal{Z}(t) = \left\{ \left[\iota_i(t), \boldsymbol{b}_i(t), \boldsymbol{m}_i(t)\right] \right\}_{i=1}^{n_t}. \tag{4.11}$$

With this computational formalization in mind, a more cognitively and developmentally motivated alternative would be to base the discretization on *motion segmentation* (Zappella et al., 2008; Anthwal & Ganotra, 2019). In such a setting, the agent's own physical interventions could help reveal object boundaries by inducing coherent object motion, reducing the need for annotated object categories and aligning more closely with active object learning[11] in infants and robots (Held & Hein, 1963; Baillargeon, 2004; Fitzpatrick et al., 2003; Anthwal & Ganotra, 2019). In the present implementation, however, segmentation is not treated as the main research problem. Moreover,

[11] For an overview of the ACV paradigm, refer to Section 1.2.2.

current state-of-the-art motion segmentation approaches remain limited in their applicability to complex real-world multi-object scenarios in terms of flexibility, robustness, and computational efficiency (Yao et al., 2020; Gao et al., 2023; Zhou et al., 2023). We therefore instantiate $\Sigma_{\boldsymbol{\theta}^{\text{seg}}}$ using a dedicated detection, segmentation, and tracking stack, which provides more reliable object hypotheses under the constraints of this work, while leaving intervention-driven motion segmentation as a future extension.

Specifically, the implemented subsystem uses the YOLOE detection and segmentation stack (Wang et al., 2025) based on YOLOv11 (Jocher & Qiu, 2024), combined with BoT-SORT-style (Aharon et al., 2022) tracking for short-term identity maintenance. The output of this stage is the stabilized object observation set $\mathcal{Z}(t)$, which is passed to the perceptual memory and downstream descriptor-construction stages. Although YOLOE uses language-conditioned open-vocabulary detection internally, this is treated here only as a technical shortcut for obtaining object hypotheses. The present architecture does not rely on natural-language grounding at this stage, and linguistic capacities are considered only as a later extension in Section 6.4.

4.2.2.2 Perceptual Memory and Maintenance

The stabilized object observations $\mathcal{Z}(t)$ obtained by the discretization process above are integrated into an explicit *perceptual working memory*. Architecturally, this memory preserves object-local observations over time, so that downstream perceptual operations do not depend solely on the current visual input. Its central component is an *ephemeral eidetic memory storage* (EEMS), which acts as an object-local reference memory, i.e., for each maintained object, it stores a small set of selected RGB-D crops, masks, and related perceptual metadata rather than the full visual stream. Formally, for each represented object i, the perceptual record can be abstractly written as

$$\omega_i(t) \triangleq \big[\iota_i(t), \Lambda_i(t), \boldsymbol{\phi}_i(t)\big], \tag{4.12}$$

where $\iota_i(t)$ is the stabilized object identifier, $\Lambda_i(t)$ is the retained set (i.e., a *gallery*) of object-local snapshots, and $\boldsymbol{\phi}_i(t)$ denotes maintenance metadata such as the last associated bounding box, visibility, or age of the object in the scene. Here, by *snapshot* we mean a selected object-local perceptual record extracted from a single RGB-D observation, containing the object crop, aligned depth support, object mask, and metadata needed for later object-level processing.

The EEMS gallery is capacity-limited. Thus, for object i,

$$\Lambda_i(t) = \{\lambda_{i,k}\}_{k=1}^{K_i(t)}, \tag{4.13}$$

where $\lambda_{i,k}$ is the k-th retained snapshot of object i, $K_i(t)$ is the current number of retained snapshots, and K_{eems} is the maximum number of snapshots retained per object, with $K_i(t) \leq K_{\text{eems}}$. In our proof-of-concept implementation (Section 5.1), a snapshot $\lambda_{i,k}$ contains an object-local RGB crop, the corresponding aligned depth crop,

an object mask, the associated bounding box, local feature information used by later pose-estimation stages (Section 4.2.2.3), when available, the observation acquisition time, and quality or diagnostic metadata.

Incoming observations are not added to $\Lambda_i(t)$ indiscriminately. The *admission mechanism* rejects low-quality object views, such as observations with too small a mask support. Among admissible views, the mechanism prefers informative reference snapshots: sharp observations are favored, with sharpness estimated using a Laplacian-based focus measure (Pech-Pacheco et al., 2000), while redundancy with already retained snapshots is assessed using image-level similarity, such as mean squared error (MSE), and local-feature overlap when available. Thus, newly retained snapshots should be both sufficiently reliable and sufficiently different from the existing gallery. Consequently, the EEMS functions as a compact object-specific visual evidence store.

This memory layer separates persistent object references from momentary detector and tracker outputs. The maintenance metadata supports short-term identity repair when tracker identifiers fragment, while the retained RGB-D snapshots and keypoint features support later pose estimation and visual descriptor construction. This mechanism should not be interpreted as a full object-permanence model: it does not explicitly infer hidden object states, predict motion through long occlusions, or perform general visual re-identification after large displacements. Rather, EEMS provides local, inspectable memory storage needed for stable object-centric active perception.

4.2.2.3 Physical Behavior Monitoring

While the EEMS preserves object-local visual evidence, the *physical behavior monitoring pipeline* is one of the three feature-extraction branches and estimates object pose and motion from successive observations. Architecturally, this stage maps maintained object observations into *kinematic descriptors* describing translation, rotation, displacement, and other interaction-relevant motion regularities. For object i, this process can be formalized abstractly as

$$\mathrm{BM}_{\boldsymbol{\theta}^{\mathrm{kin}}}:\ [\omega_i(t), \mathcal{Z}(t)] \mapsto \hat{\boldsymbol{s}}_i^{\mathrm{kin}}(t), \tag{4.14}$$

where $\mathrm{BM}_{\boldsymbol{\theta}^{\mathrm{kin}}}$ denotes the parametrized behavior-monitoring process and $\hat{\boldsymbol{s}}_i^{\mathrm{kin}}(t)$ is the estimated kinematic descriptor of object i.

In our proof-of-concept implementation, this stage estimates object pose by matching the current object observation against retained snapshots $\lambda_{i,k} \in \Lambda_i(t)$. Local image features are extracted using DeDoDe (Edstedt, Bökman, Wadenbäck, & Felsberg, 2024; Edstedt, Bökman, & Zhao, 2024) and matched using LightGlue (Lindenberger et al., 2023). Matches outside the object mask are removed, so that geometric estimation is based only on correspondences lying on the segmented object surface. The matched image points are then lifted into 3D using the aligned depth maps, producing paired reference and live point sets.

The implementation uses a hybrid geometric solver. When sufficient depth-backed correspondences are available in both the reference and current observation, pose estimation is treated as a robust 3D–3D rigid alignment problem. Formally, let $\boldsymbol{x}_{i,k,r}^{\mathrm{ref}}$, $\boldsymbol{x}_{i,r}^{\mathrm{live}}(t) \in \mathbb{R}^3$ denote the r-th matched point pair between snapshot $\lambda_{i,k}$ and the current observation. The estimated rigid transform is then obtained by minimizing

$$\left[\hat{\boldsymbol{R}}_i(t), \hat{\boldsymbol{d}}_i(t)\right] = \underset{\boldsymbol{R}\in\mathrm{SO}(3),\, \boldsymbol{d}\in\mathbb{R}^3}{\operatorname{argmin}} \sum_r \left\| \boldsymbol{x}_{i,r}^{\mathrm{live}}(t) - \left(\boldsymbol{R}\boldsymbol{x}_{i,k,r}^{\mathrm{ref}} + \boldsymbol{d}\right) \right\|_2^2, \tag{4.15}$$

where $\hat{\boldsymbol{R}}_i(t)$ and $\hat{\boldsymbol{d}}_i(t)$ denote the estimated rotation matrix and translation vector, respectively. In the implementation, this problem is approximated using RANSAC-based affine estimation (Fischler & Bolles, 1981). Since the estimated affine linear component may contain small-scale or shear artifacts, it is replaced by the closest valid rotation matrix using singular value decomposition, following the standard least-squares rigid alignment methods (Umeyama, 1991).

If the 3D–3D route is not reliable, the system falls back to a perspective-n-point (PnP) formulation, a standard pose-estimation problem from 2D–3D correspondences (Lepetit et al., 2009). In this case, reference 3D points are projected into the current image through the calibrated camera model:

$$\boldsymbol{u}_{i,r}(t) \approx \pi_{\boldsymbol{K}} \left[\hat{\boldsymbol{R}}_i(t) \boldsymbol{x}_{i,k,r}^{\mathrm{ref}} + \hat{\boldsymbol{d}}_i(t) \right], \tag{4.16}$$

where $\boldsymbol{u}_{i,r}(t)$ is the observed 2D image point corresponding to reference point $\boldsymbol{x}_{i,k,r}^{\mathrm{ref}}$, $\boldsymbol{K}$ is the camera intrinsic matrix, and $\pi_{\boldsymbol{K}}$ denotes projection into image coordinates. The PnP solver estimates the pose that best explains the observed 2D–3D correspondences. This fallback is useful when the current depth is incomplete or noisy, although it is generally less stable for weakly textured, planar, or geometrically ambiguous objects.

Having the $\hat{\boldsymbol{R}}_i$ and $\hat{\boldsymbol{d}}_i$ estimated, these are finally converted into a standard 6-degree-of-freedom (6-DoF) object pose estimation format:

$$\hat{\boldsymbol{p}}_i(t) \triangleq \left[\hat{\boldsymbol{d}}_i(t), \mathrm{rotvec}\left(\hat{\boldsymbol{R}}_i(t) \right) \right] \in \mathbb{R}^6, \tag{4.17}$$

where $\mathrm{rotvec}(\cdot)$ maps a rotation matrix to its three-dimensional rotation-vector representation. With the subsystem tracking the position of each object over the whole observation duration, the pose estimates are accumulated into object trajectories, i.e., ordered sequences of valid pose estimates recorded up to time t:

$$\mathcal{T}_i(t) \triangleq \left[\hat{\boldsymbol{p}}_i(\tau) \right]_{\tau \le t}. \tag{4.18}$$

The output of the motion estimation process is utilized in two ways: the object's immediate pose estimate is reported as a partial kinematic descriptor, i.e., $\hat{\boldsymbol{p}}_i(t) \subseteq \hat{\boldsymbol{s}}_i^{\mathrm{kin}}(t)$, and every trajectory instance $\mathcal{T}_i(t)$ is converted into a fixed-dimensional conceptual representation embeddable in a conceptual space. For the latter, we treat raw $\mathcal{T}_i(t)$ as a

signal of variable duration and sampling density. In our implementation, the trajectory signal is first preprocessed by retaining only valid pose estimates and smoothing the pose stream with a 1€ Filter (Casiez et al., 2012). The processed sequence is expressed in relative motion coordinates, so that the representation captures the shape of the movement rather than only absolute scene position. Translation is represented by frame-to-frame displacement, while orientation is converted to a rotation-vector representation to avoid discontinuities in angular coordinates. The resulting variable-length sequence is then interpolated and resampled to a fixed temporal support, after which a discrete cosine transform (DCT; Ahmed et al., 1974; Mao et al., 2019, 2020) is applied separately to the translational and rotational components. Keeping only the low-frequency coefficients yields a compact motion vector that preserves the coarse temporal structure of the trajectory while making different trajectory instances comparable within a common conceptual space representation.

It should be noted that this hybrid pipeline is chosen for pragmatic architectural reasons. Recent general-purpose 6-DoF pose-estimation and tracking systems provide stronger pose estimates for novel objects, often through render-and-compare, neural reconstruction, or foundation-model-style formulations (Labbé et al., 2022; Wen et al., 2023, 2024; Lin et al., 2024; Lee et al., 2025). However, the current subsystem does not treat 6-DoF pose tracking as its primary objective. Pose estimation is used here as an intermediate source of object-centric motion evidence. The chosen RGB-D feature-matching and geometric-estimation pipeline is therefore less general and less accurate than specialized state-of-the-art pose systems, but it is fast, transparent, modular, and sufficient for producing inspectable motion descriptors in the controlled proof-of-concept setting.

4.2.2.4 Visual Quantization

The *visual quantization branch* extracts appearance-related descriptors from object-local snapshots retained in EEMS. While physical behavior monitoring (Section 4.2.2.3) estimates how an object moves, visual quantization estimates relatively static visual properties of the object, such as shape, color, and coarse texture. For object i, this branch can be abstractly formalized as

$$\mathrm{VQ}_{\boldsymbol{\theta}^{\mathrm{app}}} : \Lambda_i(t) \mapsto \hat{\boldsymbol{s}}_i^{\mathrm{app}}(t), \tag{4.19}$$

where $\mathrm{VQ}_{\boldsymbol{\theta}^{\mathrm{app}}}$ denotes the $\boldsymbol{\theta}^{\mathrm{app}}$-parametrized appearance-quantization process and $\hat{\boldsymbol{s}}_i^{\mathrm{app}}(t)$ is the estimated appearance descriptor of object i.

In the general architecture, VQ may be instantiated by pretrained object-centric representation methods, such as slot-based models (Locatello et al., 2020) or neural concept-binder-style mechanisms (Stammer et al., 2024). In our implementation, however, the appearance descriptor is deliberately handcrafted and interpretable. This

choice avoids the need for additional representation learning from a small dataset and keeps the resulting visual dimensions inspectable.

Concretely, each retained snapshot is mapped to a vector of object-local visual features computed from the RGB crop and the object mask. The implemented descriptor includes mask geometry and Hu moment invariants for shape (Hu, 1962), HSV color histograms for coarse color distribution (Swain & Ballard, 1991), and edge density or gradient orientation statistics for contour- and texture-like structures (Dalal & Triggs, 2005). Descriptors from multiple retained snapshots can be aggregated to obtain an object-level appearance estimate:

$$\hat{\boldsymbol{s}}_i^{\text{app}}(t) = \text{agg}\left(\left\{\, \text{VQ}\left(\lambda_{i,k}\right) \right\}_{k=1}^{K_i(t)}\right), \tag{4.20}$$

where $\text{agg}(\cdot)$ denotes an aggregation operation, such as averaging across retained snapshots.

The resulting descriptor is not intended as a general object-recognition embedding. Its role is to provide a compact, interpretable projection of visual variation that can support similarity comparison, prototype construction, and later conceptual representation. Hence, visual quantization can be understood as a transformation from object-local sensory evidence to appearance-related quality dimensions in the perceptual conceptual space.

4.2.2.5 Scene Composition Parsing

The *scene composition parsing branch* is the third feature-extraction branch of the perceptual subsystem. While visual quantization (Section 4.2.2.4) and physical behavior monitoring (Section 4.2.2.3) describe individual objects' appearance and motion, scene composition parsing estimates relations between represented objects, including spatial layout, overlap, relative motion, and other interaction-relevant configurations. In including and structuring this subsystem, we follow *qualitative spatial-reasoning* (QSR) approaches (Cohn & Renz, 2008), in which metric perceptual quantities are lifted into relations such as distance, orientation, contact, overlap, and containment, with topological relations between regions commonly formalized through region-based spatial calculi (Egenhofer & Franzosa, 1991; Randell et al., 1992); this idea also aligns with scene-graph approaches (Chang et al., 2023) in computer vision, where scenes are represented as structured object–attribute–relation graphs for visual reasoning (Johnson et al., 2015; Krishna et al., 2017).

In our architecture, for a scene with n_t represented objects, the process can be formulated abstractly as

$$\text{SCP}_{\boldsymbol{\theta}^{\text{scene}}}\colon \left\{\hat{\boldsymbol{s}}_i^{\text{obj}}(t)\right\}_{i=1}^{n_t} \mapsto \hat{\boldsymbol{s}}^{\text{scene}}(t), \tag{4.21}$$

where $\text{SCP}_{\boldsymbol{\theta}^{\text{scene}}}$ denotes the scene-composition process and $\hat{\boldsymbol{s}}^{\text{scene}}(t)$ is the estimated scene-level descriptor.

At the object level, the parser estimates an egocentric spatial descriptor $\hat{\boldsymbol{s}}_i^{\text{spat}}(t)$ for each represented object. This descriptor summarizes the object's position relative to the agent or camera frame, for example, by combining an estimated 3D centroid, distance, azimuth/elevation, and image- or depth-support information. The exact fields depend on the available perceptual evidence and the downstream representation used by the WM.

At the relational level, the parser constructs pairwise descriptors between represented objects. Given estimated object 3D centroids $\boldsymbol{\mu}_i^{\text{pos}}(t), \boldsymbol{\mu}_j^{\text{pos}}(t) \in \mathbb{R}^3$, the relative displacement from object i to object j is simply $\boldsymbol{\mu}_j^{\text{pos}}(t) - \boldsymbol{\mu}_i^{\text{pos}}(t)$. Together with derived quantities such as relative distance, relative orientation, overlap, containment-like support, and relative motion, this displacement defines a pairwise relational descriptor $\hat{\boldsymbol{s}}_{ij}^{\text{rel}}(t)$. Topological and contact-like relations can be approximated from masks, bounding boxes, or depth-supported object regions using overlap criteria such as the IoU and IoM measures introduced in Section 4.2.2.1.

In the fuller architectural specification (Fig. 4.2), pairwise descriptors are further processed by a relative-position parser, a topological overlap checker, and a kinematic synchrony estimator. These mechanisms convert centroid differences into spatial relations, evaluate mask-, bounding-box-, or depth-based overlap, and compare relative velocities or motion profiles to detect coordinated displacement. Their outputs can be compiled into a lightweight *scene grammar*, in which object nodes are connected by spatial, topological, and kinematic relations.[12] The grammar can then be vectorized or otherwise serialized into a composite scene representation usable by the WM and by later conceptual representation construction.

The scene descriptor can then be viewed as the output of a structured aggregation process applied to the current object-level descriptors. Internally, this process constructs pairwise relational descriptors and serializes their spatial, topological, and kinematic structure into a scene-grammar-like representation:

$$\hat{\boldsymbol{s}}^{\text{scene}}(t) = \text{SCP}\left[\left\{\hat{\boldsymbol{s}}_i^{\text{obj}}(t)\right\}_{i=1}^{n_t}\right]. \tag{4.22}$$

The resulting descriptor exposes relational structure that is not available from isolated object descriptors alone. This is important for downstream world modeling, since many action-relevant properties, including support, containment, contact, obstruction, or coordinated displacement, are relations between objects rather than intrinsic object properties.

In our implementation, however, this subsystem is not instantiated and remains primarily an architectural specification for future work.

[12] Scene grammar in our case can be interpreted as a similar, but weaker version of full stochastic image grammars (Zhu & Mumford, 2007).

4.2.2.6 Conceptual Representation Construction

The final stage of the perceptual subsystem projects structured perceptual descriptors into domain-specific conceptual representations. Following the conceptual-space account introduced in Section 1.2.3, perceptual qualities are represented as dimensions within metric domains that support similarity-based comparison and prototype formation (Gärdenfors, 2000, 2014).

Let $J_i(t) \subseteq \mathcal{D}$ denote the perceptual domains currently available for object i. The corresponding object-specific conceptual subspace can then be formulated as

$$\mathcal{C}_i(t) \equiv \mathcal{C}_{J_i(t)} \triangleq \prod_{\delta \in J_i(t)} \delta. \tag{4.23}$$

For each available domain $\delta \in J_i(t)$, the relevant domain-specific perceptual descriptor $\hat{\boldsymbol{s}}_{i,\delta}(t)$ is mapped into a conceptual point via

$$\zeta_\delta \colon \hat{\boldsymbol{s}}_{i,\delta}(t) \mapsto \boldsymbol{c}_{i,\delta}(t) \in \delta. \tag{4.24}$$

The object-level conceptual representation is then the collection of available domain projections:

$$\boldsymbol{c}_i(t) \triangleq \left[\boldsymbol{c}_{i,\delta}(t)\right]_{\delta \in J_i(t)} \in \mathcal{C}_i(t). \tag{4.25}$$

This partial-domain formulation allows us to represent an object in only relevant domains, as, for instance, an object may have a visual projection without a reliable motion projection, or a motion projection without a scene-level relational projection.

Observed conceptual points can be retained as exemplars. When category clusters are available, a prototype-like summary for category ℓ in domain δ can be represented by

$$\boldsymbol{\mu}_{\ell,\delta} = \frac{1}{N_{\ell,\delta}} \sum_{m=1}^{N_{\ell,\delta}} \boldsymbol{c}_{\ell,\delta}^{(m)}, \tag{4.26}$$

where $N_{\ell,\delta}$ is the number of observed exemplars assigned to category ℓ in domain δ. This gives us a simple computational counterpart of the exemplar/prototype conceptualization discussed in Section 1.2.3, while avoiding the stronger assumption that concepts are already represented as full convex regions (Gärdenfors, 2000).

In our proof-of-concept implementation, conceptual construction is only partially realized. The visual branch maps retained object snapshots into a visual conceptual space using the handcrafted appearance descriptor described in Section 4.2.2.4, while the motion branch maps recorded object trajectories into a motion conceptual space using the DCT-based trajectory vectorization described in Section 4.2.2.3. The spatial and scene-grammar domains remain architectural specifications intended for future work.

Conceptual representation construction thus links perception to world modeling by converting selected visual, spatial, and kinematic descriptors into stable object-level variables for comparison, categorization, and later sensorimotor or causal learning.

4.2.3 Neurocognitive Correlates

The perceptual subsystem described above is not intended as a biologically faithful model of the visual system. Nevertheless, its organization is motivated by several neurocognitive principles introduced in Section 1.2. The associations with some neural areas highlighted in Fig. 4.3 should therefore be understood as functional analogies – they indicate which biological functions the corresponding computational mechanisms approximate, not a one-to-one mapping between modules and cortical regions. The purpose of this subsection is to make explicit which neurocognitive principles guided the subsystem's functional decomposition.

At the input level (Section 4.2.2.1), RGB-D co-alignment, segmentation, and tracking function as the architecture's analog of early visual organization. Rather than modeling the early retina–LGN–V1 pathway explicitly, the subsystem recovers low-level structure relevant to object individuation, such as image support, contours, depth discontinuities, and candidate object boundaries (Marr, 2010; Nakayama et al., 1989; Kellman & Shipley, 1991). The depth-supported and tracking-related aspects of this stage are functionally similar to dorsal visual processing, where spatial layout and motion cues contribute to locating and tracking objects for action (Goodale & Milner, 1992; O'Reilly et al., 2024). Cognitively, the same role is related to developmental accounts of object individuation, where cohesion, continuity, common motion, and interaction help reveal object unity under changing viewpoint and partial occlusion (Spelke, 1990; Baillargeon, 2004). In the current implementation, the detector-tracker stack acts as a pragmatic substitute for a more developmentally plausible motion-based mechanism, in which active intervention can expose object boundaries and physical coherence (Held & Hein, 1963; Fitzpatrick et al., 2003; Anthwal & Ganotra, 2019). At this stage, however, the output is only a set of candidate object observations, which must still be maintained and enriched by later perceptual mechanisms.

The EEMS component (Section 4.2.2.2) then provides a limited object-centered perceptual memory. Cognitively, its role is close to object-file (Kahneman et al., 1992) or visual-index accounts (Pylyshyn, 2001), in which the visual system maintains temporary object-specific references that allow object features to remain bound to the same perceived entity across short temporal intervals. This parallel is also consistent with the capacity-limited character of visual working memory (Luck & Vogel, 1997; Cowan, 1998). Neurally, the short-term retention of object-specific visual information is distributed across multiple regions: object identity and appearance are associated with ventral visual and inferotemporal representations, while attentional selection and active maintenance involve broader frontoparietal control mechanisms (Desimone & Duncan, 1995; Miller & Desimone, 1994; Xu & Chun, 2006). Accordingly, EEMS is best understood as an inspectable reference-maintenance mechanism supporting stable object-centered processing across successive observations, not as a bio-plausible visual memory or a full model of object permanence.

The subsequent organization of the feature-extraction branches follows the broad functional division between ventral and dorsal visual processing discussed in Section 1.2.1 (Goodale & Milner, 1992; O'Reilly et al., 2024). The visual quantization process (Section 4.2.2.4) is closest to the VVP, since it extracts relatively stable appearance-related properties for object-centered comparison and later conceptual representation. Physical behavior monitoring (Section 4.2.2.3), by contrast, functionally resembles the DVP, with the proposed kinematic pipeline being especially analogous to the posterior parietal cortex (PPC), which is commonly associated with transforming visual information into spatial and sensorimotor representations for action (Andersen & Buneo, 2002; Culham & Valyear, 2006).

Scene composition parsing (Section 4.2.2.5) provides the scene-level counterpart to object-centered processing. Neurocognitively, this subsystem is related to parahippocampal processing of visual scene layout, spatial context, and navigationally relevant environmental structure (Epstein & Kanwisher, 1998; Epstein, 2008; Aminoff et al., 2013). At the cognitive level, it is motivated by QSR accounts (Cohn & Renz, 2008), which treat spatial structure as relations over perceived entities rather than as a flat collection of object features (Egenhofer & Franzosa, 1991; Randell et al., 1992). Cognitive-map accounts, from Tolman (1948) to hippocampal–entorhinal theories of relational memory (Eichenbaum, 2017; Behrens et al., 2018; Whittington et al., 2020), further motivate relational structure as a scaffold for prediction and flexible inference.

Finally, conceptual representation construction (Section 4.2.2.6) corresponds to the transition from object-local perceptual measurements to reusable semantic variables. This follows grounded and distributed accounts of semantic cognition, where concepts are supported by perceptual and action-related systems rather than by amodal symbols alone (Barsalou, 1999; Lambon Ralph et al., 2016). Neurally, the ATL has been interpreted as a semantic hub that integrates modality-specific information into more abstract conceptual representations (Patterson et al., 2007; Lambon Ralph et al., 2016; Frisby et al., 2023). In the proposed architecture, conceptual spaces (Gärdenfors, 2000) provide the computational counterpart by organizing perceptual descriptors into domain-specific metric representations that remain grounded in the visual, spatial, and kinematic branches.

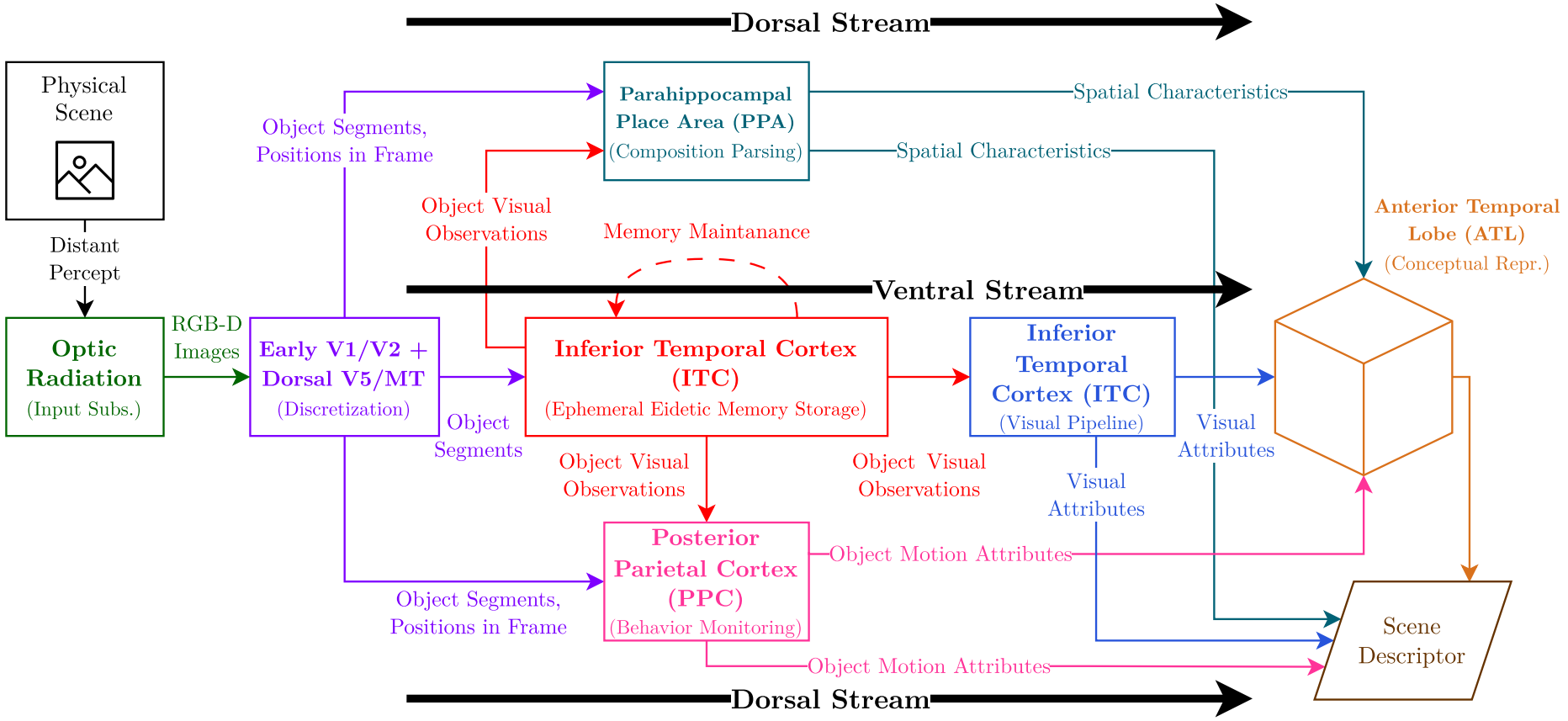


Figure 4.3: Functional neurocognitive interpretation of the proposed perceptual system. The labels indicate hypothesized approximate functional correspondences between architectural mechanisms and neurocognitive visual-processing principles, not direct claims about biological implementation.

4.3 Neurosymbolic Knowledge Representation

The WM subsystem constitutes the second main representational component of the proposed architecture. While the perceptual subsystem maps raw visual input to structured object-centric descriptors, the WM operates on the structured sensorimotor history $\mathcal{H}_t$ (Eq. 4.4). Its role is to learn and maintain an explicit model of dependencies between actions, proprioceptive variables, perceptual state descriptors, inferred latent variables, and dynamically selected target variables. The resulting model supports prediction, state inference, intrinsic target generation, and action generation within the same PAL.

The subsystem is neurosymbolic in the specific sense that its variables and dependency structure are explicit and inspectable, while its *local mechanisms* may be implemented by differentiable learned estimators.[13] The model therefore combines a *graph-level structure* providing symbolic and causal organization with local mechanisms providing subsymbolic approximations of continuous sensorimotor relationships. It should not be understood as a full transition kernel T over the complete state space $\mathcal{S}$, nor as a replacement for the perceptual mapping Π. Instead, it is a learned structured model over selected variables exposed by perception, proprioception, and internal inference. In proof-of-concept settings (Section 5.2), selected perceptual descriptors may be treated as observed inputs to the WM, but this is a simplifying assumption rather than a claim that perception recovers the true environmental state.

4.3.1 Model Structure

Following the causal formalism introduced in Section 1.3.2, we specify the WM as an SCM-like model (Pearl, 2009; Peters et al., 2017) over selected endogenous variables:

$$\mathcal{M}^{\mathrm{wm}}(t) \triangleq \left\langle \mathcal{V}^{\mathrm{wm}}(t), \mathcal{U}^{\mathrm{wm}}(t), \mathcal{F}^{\mathrm{wm}}(t), p_t\left(\mathcal{U}^{\mathrm{wm}}\right) \right\rangle, \tag{4.27}$$

where $\mathcal{V}^{\mathrm{wm}}(t)$ is the set of endogenous variables represented by the WM, $\mathcal{U}^{\mathrm{wm}}(t)$ denotes exogenous residual variables, $\mathcal{F}^{\mathrm{wm}}(t)$ denotes the set of structural mechanisms, and $p_t\left(\mathcal{U}^{\mathrm{wm}}\right)$ denotes the current residual distribution. The induced DAG $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t)$ makes the dependency structure inspectable, i.e., an edge $X_j \to X_i$ indicates that X_j is treated as a direct parent of X_i in the corresponding structural mechanism.

Furthermore, we partition the endogenous variables by their architectural role (Fig. 4.4):

$$\mathcal{V}^{\mathrm{wm}}(t) \triangleq \mathcal{V}^{\mathrm{act}} \cup \mathcal{V}^{\mathrm{obs}} \cup \mathcal{V}^{\mathrm{lat}}(t), \tag{4.28}$$

where $\mathcal{V}^{\mathrm{act}}$ contains *action or control variables*, $\mathcal{V}^{\mathrm{obs}}$ contains *variables observed* or estimated from the current sensorimotor state, and $\mathcal{V}^{\mathrm{lat}}(t)$ contains *latent variables*

[13] For an overview of the neurosymbolic paradigm and multilevel architectures as we understand them in the context of this thesis, see Section 1.3.4.

inferred internally by the model. The sets $\mathcal{V}^{\text{act}}$, $\mathcal{V}^{\text{obs}}$, and $\mathcal{V}^{\text{lat}}(t)$ are pairwise disjoint. In the simplest case, $\mathcal{V}^{\text{act}}$ corresponds to scalar components of $\boldsymbol{a}(t) \in \mathcal{A}$, while $\mathcal{V}^{\text{obs}}$ contains selected components of $\hat{\boldsymbol{s}}(t) = \hat{\boldsymbol{s}}^{\text{vis}}(t) \cup \boldsymbol{q}(t)$.

Within the $\mathcal{M}^{\text{wm}}$, we also recognize dynamically selected *target variables*, which

$$\mathcal{V}^{\text{tar}}(t) \subseteq \mathcal{V}^{\text{act}} \cup \mathcal{V}^{\text{obs}} \cup \mathcal{V}^{\text{lat}}(t). \tag{4.29}$$

The target variables are not a separate type of causal variable, but variables whose values are temporarily preferred during inference or action generation. A target may be provided externally, as in goal-directed behavior, or generated internally by the intrinsic-motivation mechanism described in Section 4.3.3. Consequently, $\mathcal{V}^{\text{tar}}(t)$ can change across interaction steps and planning episodes without changing the underlying variable type.

Action variables $\mathcal{V}^{\text{act}}$ are treated as directly manipulable control inputs. By default, they are root variables in $\mathcal{G}_{\mathcal{M}^{\text{wm}}}(t)$, i.e., for each $A_i \in \mathcal{V}^{\text{act}}$,

$$\deg^{-}(A_i) = 0. \tag{4.30}$$

Executing an action in our framework is therefore represented as a causal intervention on the corresponding action variables (Pearl, 2009; Peters et al., 2017; Hellström, 2021), for example, $\text{do}(\boldsymbol{a}(t) = \boldsymbol{a})$, subject to the embodiment constraints encoded by $\mathcal{A}$. This distinguishes action execution from merely conditioning on an observed action value, thereby allowing the WM to be queried in an intervention-sensitive manner.

For each non-root endogenous variable $X_i \in \mathcal{V}^{\text{wm}}(t)$ with $\deg^{-}(X_i) > 0$, the architecture associates a *local structural mechanism*. In the *additive-noise form* used as the abstract mechanism class,

$$X_i \triangleq f_i\left[\text{pa}\left(X_i\right)\right] + U_i, \tag{4.31}$$

where $f_i \in \mathcal{F}^{\text{wm}}(t)$ is a deterministic, generally differentiable mechanism and $U_i \in \mathcal{U}^{\text{wm}}(t)$ captures residual variation, unmodeled causes, or stochasticity (i.e., noise) (Hoyer et al., 2008; Peters et al., 2017). The differentiability of f_i is important for later action generation, since it allows candidate actions to be optimized through the learned WM (see Sections 4.3.4 and 4.3.5 for more details). The structural mechanism should be distinguished from its particular parametrization. The additive form in Eq. 4.31 uses a deterministic parent-conditioned predictor (i.e., f_i) together with an exogenous residual term (i.e., U_i). In a more general structural equation,

$$X_i \triangleq g_i\left[\text{pa}\left(X_i\right), U_i\right], \tag{4.32}$$

the map g_i may be parametrized by different estimator families, ranging from simple regressors to neural conditional generative models. However, the architectural requirement for our use case is not a particular estimator class, but that each non-root variable is assigned by a local mechanism conditioned on its parents and that residual

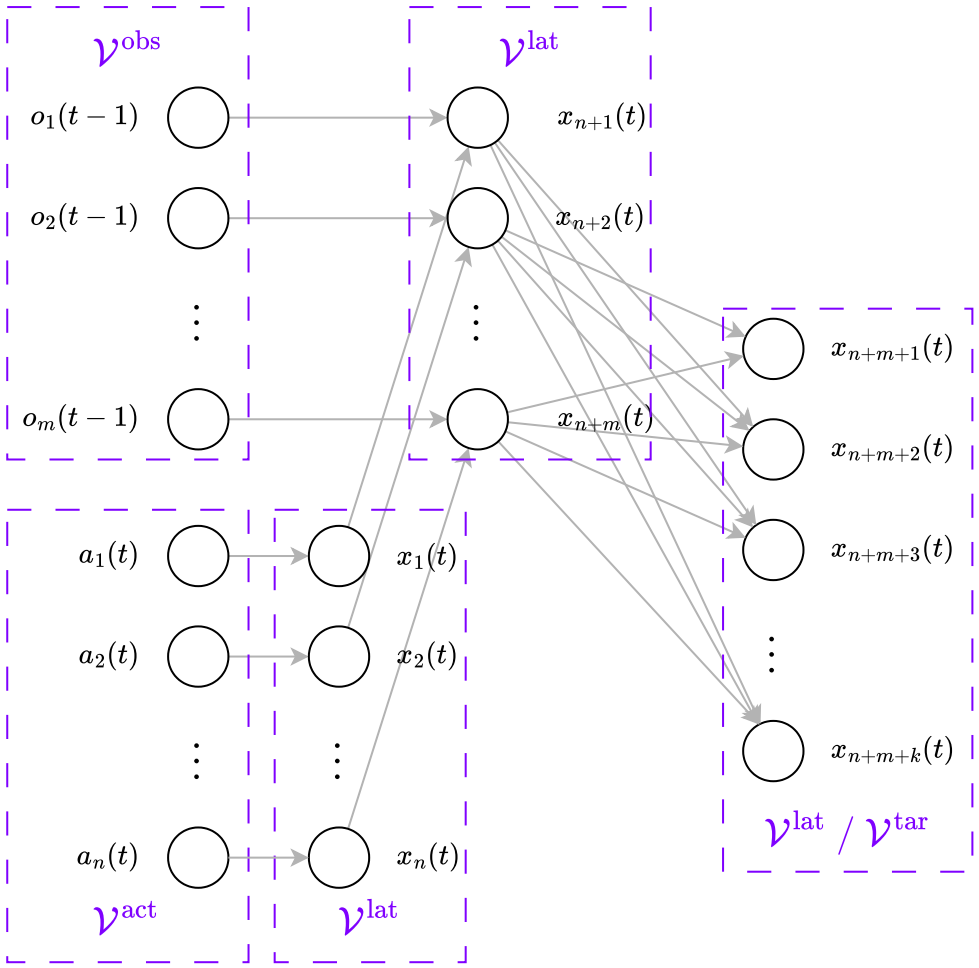


Figure 4.4: Example illustration of the WM variable types according to their topological role. Within the WM, the action variables $a_i(t) \in \mathcal{V}^{\text{act}}$ as well as the observation ones $o_i(t-1) \in \mathcal{V}^{\text{obs}}$ are represented by root nodes matching their purpose of receiving action or sensory input. Latent variables $x_i(t) \in \mathcal{V}^{\text{lat}}$ are inferred within the model, and thus are strictly non-root. Any of these variables can act as an inferential target; in the illustrated case, $x_{n+m+1}(t), \ldots, x_{n+m+k}(t) \in \mathcal{V}^{lat}$ have been selected for this role.

uncertainty remains available for prediction, state inference, and counterfactual-style abduction.

Although the notation permits vector-valued quantities for compactness, the graph $\mathcal{G}_{\mathcal{M}^{\text{wm}}}$ should be understood as operating over scalar or low-dimensional modeled variables. Thus, an action vector $\boldsymbol{a}(t)$, a proprioceptive vector $\boldsymbol{q}(t)$, or an object descriptor $\hat{\boldsymbol{s}}_i^{\text{obj}}(t)$ may be decomposed into separate graph variables when this improves causal interpretability, mechanism learning, or inference. Conversely, tightly coupled dimensions may be grouped into a single multivariate node when separating them would introduce an artificial independence assumption.

The initial graph structure can be specified in several ways. In our strongly constrained proof-of-concept setting, the graph is given by prior knowledge about the agent's embodiment, for example, by connecting action variables to proprioceptive dis-

placement variables and those variables to subsequent proprioceptive states.[14] In less constrained settings, the graph may be constructed automatically by dense initialization over admissible[15] parent relations and then sparsified through learning (Zhang et al., 2017; Vonk et al., 2023).

4.3.2 Causal Learning and Discovery

The WM is learned from the structured sensorimotor history $\mathcal{H}_t$. Each interaction step contributes evidence about relations among executed actions, current state estimates, and subsequent observed or estimated variables. In this sense, learning in $\mathcal{M}^{\mathrm{wm}}(t)$ concerns both the parametrization of local mechanisms and, when not fixed by prior knowledge, the structure of $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t)$ (Peters et al., 2017; Koller & Friedman, 2009).

For a non-root variable X_i, the relevant training evidence consists of realized parent values and the corresponding realized value of X_i extracted from $\mathcal{H}_t$. If the parent set $\mathrm{pa}(X_i)$ is fixed, mechanism learning can be formulated locally as

$$\theta_i^*(t) \triangleq \operatorname*{argmin}_{\theta_i} \sum_{\tau<t} \mathcal{L}_i \left(x_i(\tau), f_{i,\theta_i} \left[\boldsymbol{x}_{\mathrm{pa}(X_i)}(\tau) \right] \right), \tag{4.33}$$

where $\theta_i \subset \boldsymbol{\theta}^{\mathrm{wm}}$ denotes the parameters of the local mechanism for X_i, $\boldsymbol{x}_{\mathrm{pa}(X_i)}(\tau)$ denotes the realized values of its parent variables at interaction step τ, and $\mathcal{L}_i$ is an appropriate local loss, such as an MSE or negative log-likelihood. This locality is important architecturally, since each mechanism can be inspected, updated, or replaced without redefining the entire WM (Peters et al., 2017; Schölkopf et al., 2021).

In accordance with the theory introduced in Sections 1.1.2, 1.3.1 and 1.3.2, we distinguish three increasingly active data-generation regimes through which the history $\mathcal{H}_t$ can be populated:

1. *Observation-driven updating*: the model is updated from ordinary observations generated by the ongoing PAL. These data provide evidence about statistical dependencies among represented variables, but the agent does not select actions specifically to improve the WM.

2. *Randomized exploration*: the agent collects data through randomized exploratory actions, such as motor babbling (Saegusa et al., 2009; Gama, 2026). This regime broadens coverage of the action and state spaces and produces intervention-like transitions, but the actions are not selected according to any explicit estimate of expected informativeness.

[14] This model design process is referred to as causal graph initialization via expert knowledge (Pearl, 1995, 2009). See Section 5.2 for the specific implementation details.

[15] In our context, admissibility includes at least acyclicity and temporal directionality: action and state variables at the current interaction step may influence later state or sensory variables, but not conversely within the same modeled transition to not introduce anticausal relationships (Runge et al., 2019).

3. *Experimentally directed exploration*: once a preliminary WM is available, action selection can be treated as an experimental-design problem. The agent can then choose interventions expected to be informative about uncertain mechanisms, graph relations, or target variables (Toth et al., 2022; Agrawal et al., 2019).

These regimes differ in how actively the agent controls the data it receives, but all update the same structured model. The agent's own actions are especially useful because their values are selected and executed by the agent rather than only observed as naturally occurring variation. Since action execution is represented as $\mathrm{do}(\boldsymbol{a}(t) = \boldsymbol{a})$, the resulting transitions provide intervention-like information about downstream proprioceptive, perceptual, or latent variables (Pearl, 2009; Peters et al., 2017; Hauser & Bühlmann, 2012; Hellström, 2021).

When the graph structure is not fully specified a priori, learning also includes causal discovery or graph refinement. In the present architecture, this should be understood conservatively. The model may begin from an expert-specified graph (Pearl, 1995, 2009), from a dense graph over admissible parent relations, or from a partially specified structure constrained by embodiment, temporal order, and variable type. Subsequent experience can then support pruning weak or uninformative dependencies, adding plausible latent variables, or revising parent sets when persistent prediction errors indicate that the current mechanism is structurally insufficient (Pearl, 2009; Peters et al., 2017; Zhang et al., 2017; Vonk et al., 2023).

Nevertheless, note that $\mathcal{M}^{\mathrm{wm}}(t)$ should not be interpreted as discovering the complete causal structure of $\mathcal{E}$. The data available to the WM are finite and filtered through perception, while unobserved causes, sensor noise, and representation errors may still affect the learned mechanisms (as formalized in Section 3.3). Thus, $\mathcal{M}^{\mathrm{wm}}(t)$ is not assumed to recover the full causal structure of $\mathcal{E}$.

4.3.3 Intrinsic Motivation Generation

Intrinsic motivation provides the mechanism by which the WM can generate its own temporary target variables and target values (Oudeyer & Kaplan, 2007). In the present architecture, this mechanism is *epistemic*[16] rather than reward-based, since it assigns higher value to outcomes that would improve coverage of the agent's represented sensorimotor space or reduce uncertainty about the learned model (Friston et al., 2015, 2017; Parr et al., 2022). It therefore complements the externally specified targets used in goal-directed behavior.

For each target-capable variable $X \in \mathcal{V}^{\mathrm{obs}} \cup \mathcal{V}^{\mathrm{lat}}(t)$, the model maintains an estimate of the regions of its value space that have already been observed or inferred from $\mathcal{H}_t$. Let $\hat{p}_{X,t}(x)$ denote the current (i.e., at time t) *empirical density estimate* over values of X. A simple *curiosity density* can then be obtained by assigning greater probability

[16] For more details on the distinction between epistemic and pragmatic motivation, see Section 1.3.3.

to low-density regions, i.e., by inverting the explored-density estimate:

$$\tilde{p}^{\mathrm{cur}}_{X,t}(x) \triangleq \frac{1}{\hat{p}_{X,t}(x) + \varepsilon}, \tag{4.34}$$

where $\varepsilon > 0$ prevents division by zero. The normalized curiosity density then is

$$p^{\mathrm{cur}}_{X,t}(x) \triangleq \frac{\tilde{p}^{\mathrm{cur}}_{X,t}(x)}{\int_{\mathcal{X}_X} \tilde{p}^{\mathrm{cur}}_{X,t}(x')\, dx'}, \tag{4.35}$$

where $\mathcal{X}_X$ denotes the modeled support (i.e., the range) of variable X. Implementation-wise, $\hat{p}_{X,t}$ may be, in the simplest case, a histogram-based distribution; consequently, $p^{\mathrm{cur}}_{X,t}$ is obtained by assigning higher target density to sparsely visited bins.

However, purely inverse-density sampling can produce large discontinuous jumps in target space, which may require abrupt or inefficient exploratory actions. Hence, we additionally include a *locality bias* around the current value $X = x(t)$. For a scalar target variable, this mechanism can be expressed as a Gaussian modulation term

$$\Psi_X\left[x \mid x(t)\right] \triangleq 1 + \alpha \exp\left[-\frac{1}{2}\left(\frac{x - x(t)}{\sigma}\right)^2\right], \tag{4.36}$$

where $\alpha \geq 0$ controls the amplification of the locality preference and $\sigma > 0$ controls its width. The resulting modulated target density is

$$p^{\mathrm{tar}}_{X,t}[x \mid x(t)] \triangleq \frac{p^{\mathrm{cur}}_{X,t}(x)\, \Psi_X[x \mid x(t)]}{\int_{\mathcal{X}_X} p^{\mathrm{cur}}_{X,t}(x')\, \Psi_X[x' \mid x(t)]\; dx'}. \tag{4.37}$$

The locality term Ψ_X does not define the epistemic objective itself; it only biases target selection toward regions that are plausibly reachable from the current state.

A curiosity target value is then sampled from $p^{\mathrm{tar}}_{X,t}$ for a selected target variable $X \in \mathcal{V}^{\mathrm{tar}}(t)$:

$$x^*(t) \sim p^{\mathrm{tar}}_{X,t}[\,\cdot \mid x(t)]\,. \tag{4.38}$$

The action-generation subsystem (Section 4.3.5) can subsequently query $\mathcal{M}^{\mathrm{wm}}(t)$ for actions expected to realize this target. In this way, intrinsic motivation defines temporary preferred outcomes over represented variables, which are then translated into actions through the learned WM. This links our mechanism to BED and active inference: target generation specifies what would be informative to observe, while action generation determines how such an observation may be brought about (Toth et al., 2022; Parr et al., 2022).

4.3.4 State Inference

By *state inference* in this context, we understand querying $\mathcal{M}^{\mathrm{wm}}(t)$ for represented variables that are not directly available from the current structured state estimate. In contrast to the perceptual subsystem, which estimates $\hat{\boldsymbol{s}}^{\mathrm{vis}}(t)$ from $\boldsymbol{v}(t)$, WM inference

Algorithm 4.1: World-model forward query as mental simulation in the simplest residual treatment. A hypothetical action sequence is evaluated by intervening on action variables, propagating their consequences through the structural mechanisms of $\mathcal{M}^{\mathrm{wm}}(t)$, and feeding the predicted assignment forward for the next simulated step. Residuals are set to their expected values; under centered additive noise, this reduces to the deterministic mechanism output.

Input: World model $\mathcal{M}^{\mathrm{wm}}(t)$, current structured assignment $\hat{\boldsymbol{x}}(t)$, hypothetical action sequence $[\tilde{\boldsymbol{a}}(t), \ldots, \tilde{\boldsymbol{a}}(t+h'-1)]$, planning horizon h'

Output: Simulated trajectory over represented variables $\mathcal{T}^{\mathrm{sim}}_{t:t+h'}$

$\mathcal{T}^{\mathrm{sim}}_{t:t+h'} \leftarrow [\hat{\boldsymbol{x}}(t)]$
for $\tau = 0, \ldots, h'-1$ **do**
  $\hat{\boldsymbol{x}} \leftarrow \hat{\boldsymbol{x}}(t+\tau)$
  Apply intervention $\mathrm{do}[\boldsymbol{a}(t+\tau) = \tilde{\boldsymbol{a}}(t+\tau)]$ by assigning the action variables in $\hat{\boldsymbol{x}}$
  foreach $X_i \in \mathcal{V}^{\mathrm{wm}}(t)$ *in topological order of* $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t)$ **do**
    if X_i *is already assigned or intervened on* **then**
      Keep its assigned value
    else
      $\hat{x}_i \leftarrow f_i\big[\hat{\boldsymbol{x}}_{\mathrm{pa}(X_i)}\big] + \mathbb{E}[U_i]$
      Add $\hat{x}_i$ to $\hat{\boldsymbol{x}}$
  $\hat{\boldsymbol{x}}(t+\tau+1) \leftarrow$ next-step assignment extracted from $\hat{\boldsymbol{x}}$
  Append $\hat{\boldsymbol{x}}(t+\tau+1)$ to $\mathcal{T}^{\mathrm{sim}}_{t:t+h'}$

operates over the variables in $\mathcal{V}^{\mathrm{wm}}(t)$ and their learned mechanisms. Its inputs are assignments to currently available observed variables, typically selected components of $\hat{\boldsymbol{s}}(t)$, together with possible action interventions or target constraints.

For prediction, the model is evaluated forward through $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t)$. Given an intervention $\mathrm{do}(\boldsymbol{a}(t) = \tilde{\boldsymbol{a}})$ and current observed assignments, variables are evaluated in a topological order, so that parent values are available before child mechanisms are applied. For each non-intervened non-root variable X_i with an additive mechanism (Eq. 4.31),

$$\hat{x}_i = f_i\big[\hat{\boldsymbol{x}}_{\mathrm{pa}(X_i)}\big] + u_i, \tag{4.39}$$

where $\hat{\boldsymbol{x}}_{\mathrm{pa}(X_i)}$ contains the parent values available in the current forward pass: observed variables are supplied by the current state estimate, intervened action variables are fixed by $\mathrm{do}(\boldsymbol{a}(t) = \tilde{\boldsymbol{a}})$, and predicted parent variables are outputs of earlier mechanisms in the same topological evaluation. This prediction is local to the variables represented

by $\mathcal{M}^{\mathrm{wm}}(t)$; it is not a full prediction of $\boldsymbol{s}(t+1)$ under the environment transition kernel T.

This forward evaluation also provides a computational counterpart of *mental simulation* (Klein & Crandall, 1995): the agent can query its learned sensorimotor model under hypothetical action interventions and estimate their likely consequences without immediately executing them in $\mathcal{E}$ (Cibula, 2024; Cibula et al., 2024). Let $\hat{\boldsymbol{x}}(t)$ denote the current assignment over represented WM variables induced by $\hat{\boldsymbol{s}}(t)$ and any currently fixed or inferred variables. Computationally, in our case, mental simulation corresponds to repeatedly applying the *forward query* over virtual future steps: after a hypothetical action value is fixed by intervention and downstream variables are predicted, the resulting assignment can be fed back as the initial assignment for the next simulated step. Repeating this process yields a simulated trajectory

$$\mathcal{T}^{\mathrm{sim}}_{t:t+h'} = [\hat{\boldsymbol{x}}(t+\tau)]_{\tau=0}^{h'} \tag{4.40}$$

over represented variables, with $h' \leq h$ being a short planning horizon. For the formalization of this process, see Algorithm 4.1.

When latent or missing variables are represented explicitly, state inference can also be formulated in active-inference terms as approximate posterior inference. Let $\boldsymbol{x}^{\mathrm{obs}}(t)$ denote the currently available assignments to variables in $\mathcal{V}^{\mathrm{obs}}$, and let $\boldsymbol{x}^{\mathrm{lat}}(t)$ denote assignments to latent variables in $\mathcal{V}^{\mathrm{lat}}(t)$. The model may maintain an approximate posterior $q_t(\boldsymbol{x}^{\mathrm{lat}})$ and update it by minimizing VFE:

$$q_t^* \triangleq \operatorname*{argmin}_{q_t} F\big[q_t, \boldsymbol{x}^{\mathrm{obs}}(t)\big]\,, \tag{4.41}$$

where, analogously to Eq. 1.20,

$$F\big[q_t, \boldsymbol{x}^{\mathrm{obs}}(t)\big] \triangleq \mathbb{E}_{q_t(\boldsymbol{x}^{\mathrm{lat}})}\left[\log q_t\left(\boldsymbol{x}^{\mathrm{lat}}\right) - \log p_{\mathcal{M}^{\mathrm{wm}}}\left(\boldsymbol{x}^{\mathrm{obs}}(t), \boldsymbol{x}^{\mathrm{lat}}\right)\right]. \tag{4.42}$$

Here, $p_{\mathcal{M}^{\mathrm{wm}}}$ denotes the joint density over represented observed and latent variables induced by the current WM. In this formulation, state inference corresponds to finding latent assignments that best explain the currently available structured observations under the WM (Friston et al., 2015, 2017; Parr et al., 2022). Our implementation may instantiate this general principle only in simplified forms, such as forward prediction, deterministic reconstruction, or residual abduction.

Under the additive-noise assumption (Hoyer et al., 2008), the WM can also perform *residual abduction* directly. Given an observed value $x_i(t)$ and parent assignment $\boldsymbol{x}_{\mathrm{pa}(X_i)}(t)$, the residual compatible with the observation is

$$u_i(t) = x_i(t) - f_i\big[\boldsymbol{x}_{\mathrm{pa}(X_i)}(t)\big]\,. \tag{4.43}$$

This residual captures the part of the observed value not explained by the deterministic mechanism. Reusing $u_i(t)$ during short-horizon prediction conditions the simulated

rollout on the current unmodeled context instead of sampling a fresh residual at every step. In causal terminology, this corresponds to the abduction step of counterfactual reasoning, although here we apply it only within the learned approximate model (Pearl, 2009; Pearl et al., 2016).

For later action generation (Section 4.3.5), we denote the target-relevant forward query by

$$\Phi_t\colon\ [\hat{\boldsymbol{s}}(t), \boldsymbol{a}(t), \boldsymbol{u}] \mapsto \hat{\boldsymbol{x}}^{\mathrm{tar}}, \tag{4.44}$$

where Φ_t denotes the forward pass through the relevant subgraph of $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t)$ from the current structured state estimate and candidate action to the predicted target variables. The residual vector $\boldsymbol{u}$ may contain expected residuals, sampled residuals, or residuals obtained by abduction (Eq. 4.43), and $\hat{\boldsymbol{x}}^{\mathrm{tar}}$ contains the predicted values of the currently selected variables in $\mathcal{V}^{\mathrm{tar}}(t)$. Thus, state inference supplies the predictive query used by action generation to compare possible actions against externally specified or intrinsically generated targets.

4.3.5 Action Generation

Action generation treats the target variables selected in $\mathcal{V}^{\mathrm{tar}}(t)$ as preferred outcomes under the WM. We represent these preferences by a time-dependent density over target-variable values,

$$p_t\big(\boldsymbol{x}^{\mathrm{tar}} \mid C_t\big)\,, \tag{4.45}$$

where C_t denotes the current preference specification: an external goal may concentrate this density around a desired value, while an intrinsic target may use the curiosity density from Section 4.3.3.

Using the target-relevant forward query Φ_t (Eq. 4.44), we define a reduced active-inference-style *action objective* as the negative log-preference of the predicted target-relevant consequence:

$$\mathcal{L}_t^{\mathrm{act}}[\tilde{\boldsymbol{a}}(t)] \triangleq -\log p_t(\Phi_t[\hat{\boldsymbol{s}}(t), \tilde{\boldsymbol{a}}(t), \boldsymbol{u}] \mid C_t)\,. \tag{4.46}$$

This loss should not be confused with the full EFE formulation $G(\pi)$ from Eq. 1.22. $\mathcal{L}_t^{\mathrm{act}}$ is intended as a reduced preference-matching objective over represented target variables. In the intrinsic-exploration setting, C_t is generated by the curiosity mechanism, so $\mathcal{L}_t^{\mathrm{act}}$ acts as an epistemic target-reaching loss; in a goal-directed setting, C_t may encode pragmatic preferences over desired outcomes.

If the mechanisms along the relevant path are differentiable, the candidate action can be optimized by *gradient descent* (GD; Goodfellow et al., 2016, Chapter 8) through the WM while keeping $\boldsymbol{\theta}^{\mathrm{wm}}$ fixed:

$$\tilde{\boldsymbol{a}}^{(k+1)}(t) = \tilde{\boldsymbol{a}}^{(k)}(t) - \eta_{\mathrm{act}} \nabla_{\tilde{\boldsymbol{a}}^{(k)}(t)} \mathcal{L}_t^{\mathrm{act}}\big[\tilde{\boldsymbol{a}}^{(k)}(t)\big]\,, \tag{4.47}$$

where k denotes the optimization iteration and $\eta_{\text{act}} > 0$ is the action-optimization learning rate. The resulting action is afterward constrained to $\mathcal{A}$ and executed in the environment as an intervention

$$\text{do}[\boldsymbol{a}(t) = \boldsymbol{a}^*(t)]\,, \tag{4.48}$$

where

$$\boldsymbol{a}^*(t) \approx \underset{\tilde{\boldsymbol{a}}(t)\in\mathcal{A}}{\text{argmin}}\, \mathcal{L}_t^{\text{act}}[\tilde{\boldsymbol{a}}(t)]\,. \tag{4.49}$$

The exact formalization can be seen in Algorithm 4.2.

This formulation avoids requiring a separately learned direct IM (Wolpert & Kawato, 1998; Nguyen-Tuong & Peters, 2011; Baranes & Oudeyer, 2013). Rather than learning a fixed mapping from target states to actions, the system searches for actions whose predicted effects satisfy the current target query. This approach loosely follows the distal-learning idea of using a forward model to solve an inverse action problem indirectly (Jordan & Rumelhart, 1992).

Algorithm 4.2: Gradient-based action generation through the world model. The WM parameters $\boldsymbol{\theta}^{\text{wm}}$ remain fixed; optimization is performed only over the candidate action. In the simplest residual treatment, residuals are fixed to their expected values, i.e., $\bar{\boldsymbol{u}} \triangleq \mathbb{E}[\boldsymbol{U}]$.

Input: World model $\mathcal{M}^{\text{wm}}(t)$, current state estimate $\hat{\boldsymbol{s}}(t)$, preference density $p_t(\boldsymbol{x}^{\text{tar}} \mid C_t)$, expected residuals $\bar{\boldsymbol{u}}$, initial candidate action $\tilde{\boldsymbol{a}}^{(0)}(t)$, learning rate η_{act}, number of optimization steps K_{act}

Output: Optimized action $\boldsymbol{a}^*(t)$

for $k = 0, \ldots, K_{\text{act}} - 1$ **do**
- $\hat{\boldsymbol{x}}^{\text{tar},(k)} \leftarrow \Phi_t\big[\hat{\boldsymbol{s}}(t), \tilde{\boldsymbol{a}}^{(k)}(t), \bar{\boldsymbol{u}}\big]$
- $L^{(k)} \leftarrow -\log p_t\big(\hat{\boldsymbol{x}}^{\text{tar},(k)} \mid C_t\big)$
- $\tilde{\boldsymbol{a}}^{(k+1)}(t) \leftarrow \tilde{\boldsymbol{a}}^{(k)}(t) - \eta_{\text{act}} \nabla_{\tilde{\boldsymbol{a}}^{(k)}(t)} L^{(k)}$
- Enforce $\tilde{\boldsymbol{a}}^{(k+1)}(t) \in \mathcal{A}$

$\boldsymbol{a}^*(t) \leftarrow \tilde{\boldsymbol{a}}^{(K_{\text{act}})}(t)$

4.3.6 Meta-Learning

In the proposed architecture, *meta-learning* denotes structural and parametric adaptation of $\mathcal{M}^{\mathrm{wm}}$ in response to persistent inadequacies of the current WM. While ordinary mechanism learning updates the parameters θ_i of a fixed local mechanism, meta-learning concerns changes to the model configuration itself, for example, by increasing mechanism capacity, revising parent sets, introducing latent variables, splitting overly broad variables or mechanisms, or pruning dependencies that remain weak after sufficient experience.

The trigger for such adaptation is sustained prediction error or uncertainty that cannot be reduced by ordinary local mechanism fitting. For a variable X_i, this can be monitored through the residuals introduced in Eq. 4.43, for example, by tracking whether

$$\frac{1}{W} \sum_{\tau=t-W}^{t-1} \|u_i(\tau)\|^2 \tag{4.50}$$

remains high over a recent window of length W after repeated mechanism learning updates. Persistent residual structure suggests that the current parent set, mechanism class, or variable decomposition is insufficient.

This component should be understood as an architectural capacity rather than a fully autonomous self-repair procedure. In the proof-of-concept implementation, only selected forms of such adaptation may be instantiated, such as local refitting, graph pruning, or manual graph revision. Its role in the general architecture is to prevent the WM from being treated as static: as the agent explores new regions of its sensorimotor space, the encountered data distribution may shift (Schölkopf et al., 2021), and the model may need to consolidate reliable mechanisms and revise those that no longer explain the accumulated history $\mathcal{H}_t$.

Chapter 5

Experiments and Results

This chapter reports proof-of-concept experiments evaluating selected implemented components of the proposed architecture. The experiments are divided into two blocks. The first (Section 5.1) evaluates whether the perceptual subsystem can transform recorded RGB-D object observations into object-centric records, conceptual projections, multidomain entities, and scene descriptors. The second (Section 5.2) evaluates whether a minimal structured WM can learn proprioceptive action-effect regularities and support prediction, target-conditioned action inference, and curiosity-conditioned target selection. The experiments are interpreted as tests of representational feasibility rather than as benchmarks for general visual perception, causal discovery, or autonomous robotic control.

5.1 Perceptual Results

This section evaluates the implemented proof-of-concept components of the perceptual subsystem described in Section 4.2. The experiments are not intended as benchmark comparisons against specialized computer-vision systems. Instead, they examine whether the implemented pipeline can transform RGB-D observations into the object-centric and conceptual representations required by the proposed architecture. Consequently, the results are interpreted in terms of representational feasibility, diagnostic transparency, and architectural limitations rather than task-level performance. The proof-of-concept implementation of this part can be found in Appendix A.1.

5.1.1 Implementation Overview

The following experiments evaluate the implemented subset of the perceptual subsystem specified in Section 4.2.2, i.e., RGB-D object detection, segmentation and tracking, object-local EEMS snapshots, snapshot-based pose estimation, pose filtering, trajectory export, visual descriptor extraction, and conceptual-space construction. The scene-composition parsing pipeline, including explicit pairwise spatial relations and

Table 5.1: Objects and experimental variants used in the perceptual experiments: Exp. A1 inspected motion tracking, A2 used available visual descriptors, and A3 constructed multidomain entities only from reliable motion projections and available visual projections. Here, *depth translation* denotes the translation of an object towards the camera.

Object	Variant	Usable trajectory	Visual descriptor	Used in	Notes
box	translation	yes	yes	A1, A2, A3	reliable motion exemplar
box	depth translation	yes	yes	A1, A2, A3	reliable depth-motion exemplar
box	rotation	no	no	A1	inspected as rotation failure case
boat	translation	no	yes	A1, A2, A3	trajectory inspected, but excluded from motion projection
boat	rotation	no	yes	A1, A2, A3	trajectory inspected, but excluded from motion projection
ball	roll	no	yes	A1, A2, A3	no usable 6-DoF trajectory

scene-grammar construction, was not part of the evaluated implementation.

The evaluated data were obtained from saved RGB-D recordings of object interactions collected using the perceptual implementation. The recordings were captured with an RGB-D stereo camera mounted in an oblique overhead configuration, viewing the workspace from above and in front, as shown in Fig. 5.1a. The evaluated object set (Fig. 5.1b) consisted of a box with regular geometry, a boat model with irregular geometry and an asymmetrical visual structure, and a rubber ball lacking stable local texture features. These objects were manipulated during several short sessions, producing the interaction variants summarized in Table 5.1. At runtime, the scene maintains persistent object records, each containing retained RGB-D snapshots, the latest object support, a pose trajectory, and pose-estimation diagnostics. Pose estimation uses the feature-matching and RGB-D geometric procedure described in Section 4.2.2.3.

The kinematic experiment (Section 5.1.2) uses exported trajectory artifacts rather than the live scene objects directly. Each selected object trajectory is exported along with metadata containing vectorizer settings, pose quality summaries, and estimator diagnostics. Valid trajectories are then embedded into a fixed-dimensional motion conceptual space.

The visual experiment (Section 5.1.3) uses the retained object snapshots from saved scenes. Each object record is mapped to a visual conceptual point by averaging snapshot-level descriptors. Depth is not included in this visual descriptor, since the visual experiment is intended to represent appearance rather than spatial position or surface structure.

Finally, the multidomain experiment (Section 5.1.4) combines the single-space con-

Table 5.2: Configuration of the perceptual representations evaluated in the experiments.

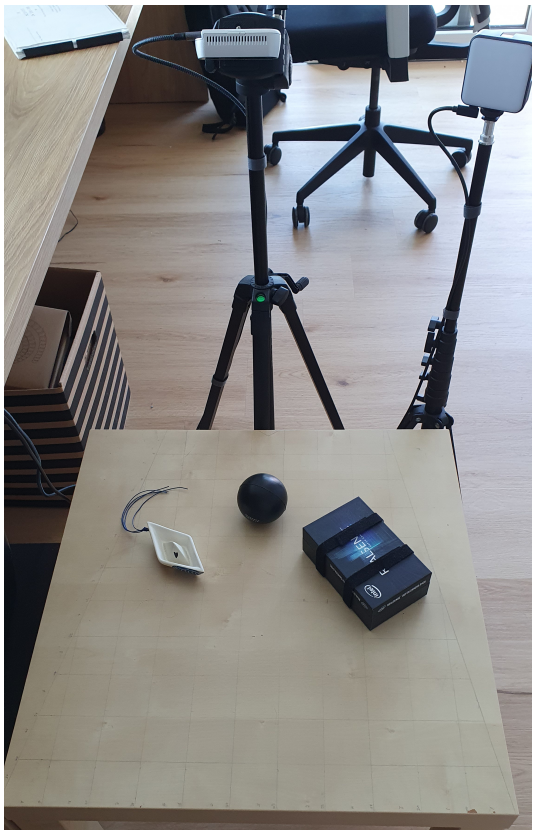
Component	Configuration	Evaluated output
Kinematic vectorization	Relative translation and rotation resampled to 60 samples; 15 low-frequency DCT coefficients retained for each of the three translational and three rotational components.	90-dimensional motion point
Visual quantization	6 mask-geometry features, 7 Hu moments, 8×4 HSV histogram bins, 1 edge-density feature, and 8 gradient-orientation bins.	54-dimensional visual point
Entity representation	Motion projections retained only for runs with usable pose trajectories; visual projections retained for all selected object records.	5 entities over motion and visual spaces

ceptual representations into entity-level representations. A conceptual entity may have a motion projection, a visual projection, or both. When a perceptual domain is not supported by reliable evidence, the entity remains undefined in that domain, and comparisons are restricted to the domains that are actually available, as proposed in Section 4.2.2.6.

The concrete representation configurations used by the experiments are listed in Table 5.2. The results should be interpreted as an evaluation of representational feasibility. They test whether the implementation can produce inspectable object-centric trajectories, visual conceptual points, and partial multidomain entities from recorded RGB-D observations. They do not evaluate the full theoretical perceptual architecture, nor do they constitute a benchmark for segmentation, visual recognition, or general 6-DoF pose tracking.

5.1.2 Experiment A1: Kinematic Concept Formation

The first experiment evaluated whether the perceptual system can recover object trajectories that can be embedded into a motion conceptual space. The analysis used recorded object-motion sequences covering successful translational motion, partial tracking, and failure cases (Table 5.1). Thus, the experiment characterizes both the conditions under which the kinematic branch produces usable motion descriptors and the conditions under which the current pose-estimation subsystem fails to provide



(a) Experimental setup with an RGB-D stereo camera mounted in an oblique overhead configuration, viewing the workspace from above and in front.

(b) Object set used in the experiments, along with the full workspace area captured from the camera's point of view.

Figure 5.1: Real-world experimental setup for the perceptual experiments A1–A3.

reliable[17] geometric trajectories.

The trajectory plots in Fig. 5.2 show the four exported records used for the motion-space projection (Table 5.3): box translation, box depth translation, boat translation, and boat rotation. Of these, the two translational box recordings provide the clearest positive result. In both cases, the object remained sufficiently visible, and feature matching remained stable enough to recover a coherent object trajectory. The exported trajectories could be transformed into fixed-dimensional motion vectors in $\mathbb{R}^{90}$, stored as conceptual points, compared by distance, and summarized by simple prototype centroids.

The remaining motion variants delimit this result. Rotation tracking was unreliable for both the box and the boat object. For example, one box-rotation recording contained 98 valid trajectory poses out of 131, but only 11 successful geometric pose

[17]In this experiment, a trajectory is treated as *reliable* only when most of its motion is supported by successful geometric pose estimates. Valid trajectory samples may also include poses propagated from image-space tracking when geometric pose estimation fails. Such trajectories are still useful for inspecting the pipeline, but they are not treated as reliable motion-space exemplars.

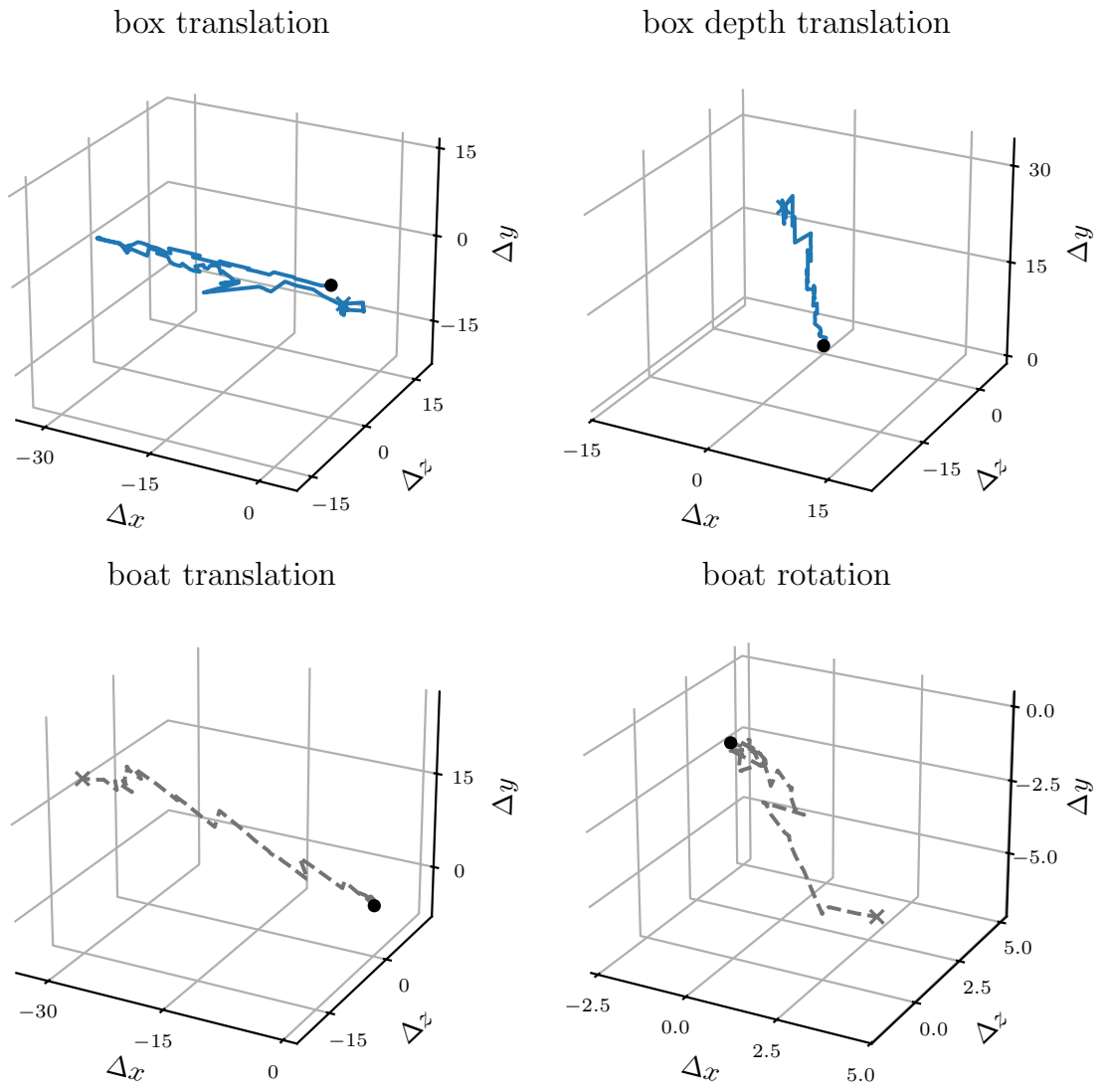


Figure 5.2: Relative 3D object trajectories exported from the motion-tracking pipeline. The coordinates are expressed in centimeters in the camera coordinate frame, with Δx denoting lateral displacement, Δz depth displacement, and Δy vertical displacement. *Black circles* and *crosses* mark trajectory starts and ends, respectively. *Solid blue trajectories* denote reliable exports, while dashed grey ones denote partial or limitation cases.

estimates from 44 estimator attempts. The remaining valid poses were mostly propagated from bounding-box displacement and therefore should not be interpreted as newly estimated orientations. The main failure mode was insufficient stable feature matches after the visible object's appearance departed from the stored reference view.

Ball rolling did not yield a usable 6-DoF pose trajectory. This is consistent with the estimator's design: a smooth or symmetric object yields weak local features, and its orientation is geometrically ambiguous for rigid-pose estimation. The segmentation

Table 5.3: Motion trajectory quality summary. T denotes recording duration, *Snap.* the number of retained object snapshots, and *Valid/total* the number of valid trajectory poses over all recorded pose entries. *Est./attempts* counts successful geometric pose estimates over estimator attempts. *Matches* denotes the mean number of feature matches retained after object-mask filtering. Only the two translational box trajectories were treated as reliable motion-space exemplars; the remaining rows indicate limitation cases.

Object	Motion	T [s]	Snap.	Valid/total	Est./attempts	Matches
box	translation	54.4	2	271/272	29/30	89.2
box	depth translation	34.6	2	199/199	19/19	91.3
box	rotation	16.5	1	98/131	11/44	27.5
boat	translation	39.0	6	214/219	20/25	31.0
boat	rotation	32.8	7	192/201	15/24	16.9
ball	roll	–	–	–	–	–

and tracking stages may still detect such an object, but the current pose-estimation subsystem is not appropriate for representing its motion as a full 6-DoF trajectory.

The resulting motion-space visualization (Fig. 5.3) should therefore be interpreted cautiously. The experiment included four trajectory vectors in $\mathbb{R}^{90}$, but only two of them were reliable pose examples. The corresponding PCA projection is useful as a diagnostic visualization showing that the exported trajectories can be embedded and inspected. However, it is not evidence of stable motion-concept clusters. The motion prototypes had support 2 for translation, support 1 for depth translation, and support 1 for rotation; the single-support prototypes are therefore stored exemplar references rather than robust category summaries.

5.1.3 Experiment A2: Visual Concept Formation

The second experiment evaluated whether object-local snapshots retained by EEMS can be embedded into a visual conceptual space independently of successful motion tracking. Five visual records were loaded from saved scene data, summarized in Table 5.4. Each record was represented by averaging the 54-dimensional snapshot descriptors available for the selected object.

This experiment produced five visual conceptual points in $\mathbb{R}^{54}$. The ball was included as a visual-only exemplar, since it had an appearance descriptor but no usable motion trajectory. The exact distances to label prototypes are reported in Table 5.4.

The visual distances should be treated as descriptive diagnostics rather than classification scores. Prototype support is low, with only two box exemplars, two boat exemplars, and one ball exemplar. The descriptor also does not produce clean object-label separation in all cases; for instance, the box depth-translation record is closer to

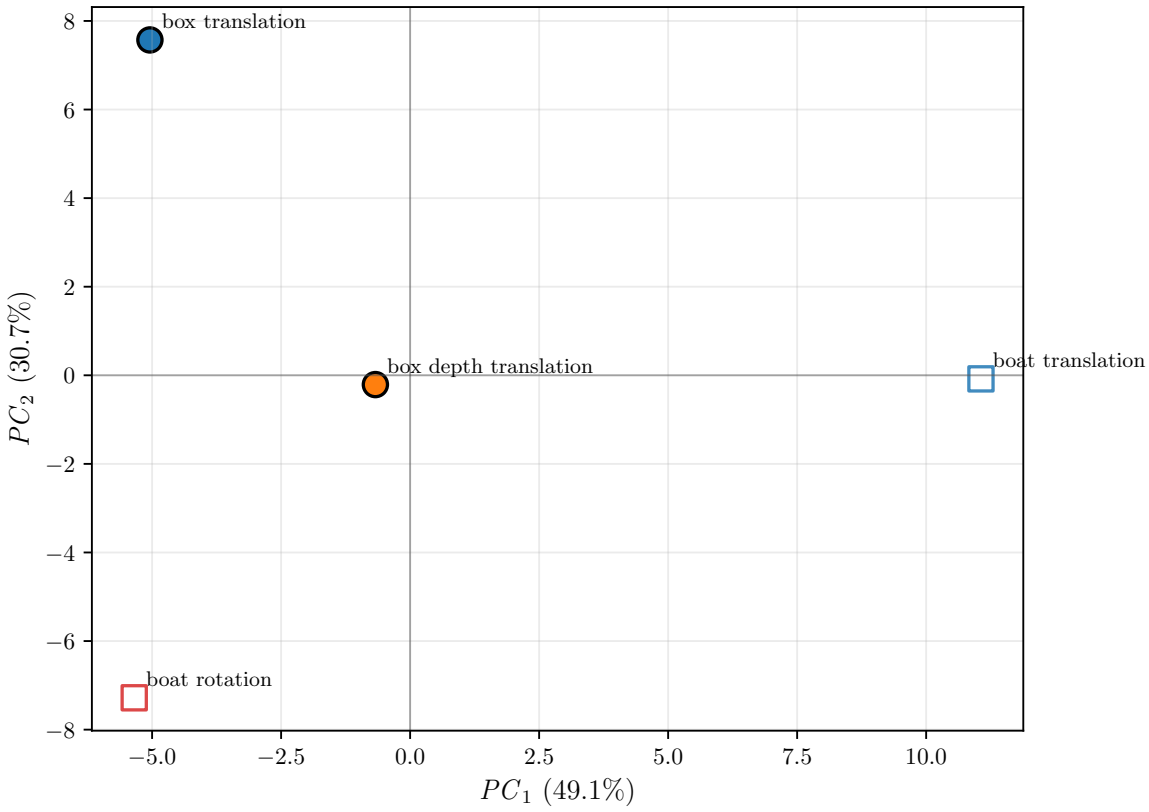


Figure 5.3: Motion conceptual-space projection of exported trajectories. PCA projection $\mathbb{R}^{90} \to \mathbb{R}^2$ of the exported motion descriptors. *Axis percentages* denote explained variance. *Filled markers* denote reliable motion exemplars; *hollow* markers denote trajectories retained as limitation cases.

the single-exemplar ball prototype than to the box prototype. This is consistent with the handcrafted descriptor's sensitivity to silhouette, viewpoint, segmentation quality, and visible surface structure.

5.1.4 Experiment A3: Multidomain Representation

The third experiment evaluated the entity-level conceptual abstraction. Rather than concatenating motion and visual vectors into a single fused representation, each object/run is represented as a conceptual entity with optional projections into named conceptual spaces. If a domain is not supported by reliable perceptual evidence, the corresponding projection is left unavailable.

The experiment had access to four exported motion descriptors in $\mathbb{R}^{90}$ and five visual-space points in $\mathbb{R}^{54}$. For the entity-level representation, however, motion projections were retained only for the two reliable box trajectories. As a result, the final entity set contained five entities, shown in Table 5.5.

The resulting projection coverage was two entities with both motion and visual projections and three entities with only a visual projection. Shared-space comparison,

Table 5.4: Visual records and distances to label prototypes in the 54-dimensional visual conceptual space, where lower values indicate greater similarity. Prototype supports were 2 for the box, 2 for the boat, and 1 for the ball. Hence, the ball prototype is identical to its single exemplar.

Object	Associated motion	Usable motion	Snapshots	d_{box}	d_{boat}	d_{ball}
box	translation	yes	2	8.778	20.000	18.240
box	depth translation	yes	2	8.778	8.831	4.937
boat	translation	no	6	11.670	3.979	5.239
boat	rotation	no	7	14.810	3.979	8.920
ball	roll	no	8	10.074	6.138	0.000

Table 5.5: Conceptual entities constructed in Exp. A3. Two entities had both motion and visual projections, while three had only visual projections. Strict comparison requiring both spaces was therefore possible only for the two box records.

Object	Run	Motion proj.	Visual proj.	Strict motion-visual dist.
box	translation	yes	yes	9.31 to box depth translation
box	depth translation	yes	yes	9.31 to box translation
boat	translation	no	yes	–
boat	rotation	no	yes	–
ball	roll	no	yes	–

therefore, allowed all entities to be compared where a common projection existed, while strict joint comparison requiring both motion and visual spaces produced only a single comparable pair: box depth translation and box translation, with distance $d = 9.31$.

The distance values are reported only within their comparison context. A visual-only distance and a motion-visual distance are not directly equivalent because they are computed over different shared spaces. See Section 1.2.3 for a detailed reasoning.

5.1.5 Structured Scene Descriptor

In addition to the three perceptual experiments described above, the implemented system contains a real-time scene descriptor processor. Its role is to expose the current state of the object-centric perceptual pipeline in a structured form usable by downstream components. The descriptor is read-only with respect to the scene; i.e., it does not perform segmentation, pose estimation, or conceptual extraction, but it serializes the currently maintained scene state and exposes object identities, image geometry, kinematic state, retained evidence, and diagnostics in a structured format. Pairwise spatial relations, topological relations, and scene-grammar construction were not included in this component.

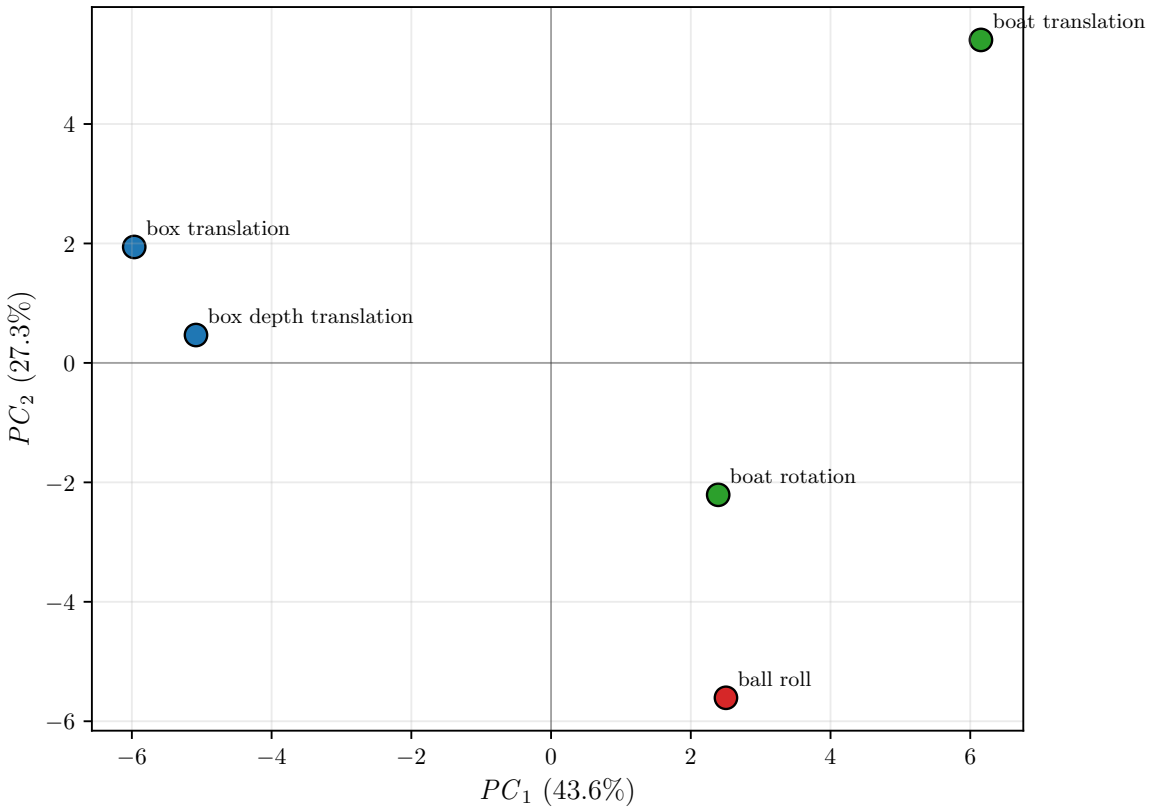


Figure 5.4: PCA projection $\mathbb{R}^{54} \to \mathbb{R}^2$ of the visual conceptual-space points. Points of the same color refer to the same physical object. The *percentages* denote the fraction of variance explained by the corresponding principal component.

5.2 World Model Results

Similar to Section 5.1, this section reports the proof-of-concept results for the WM component (Section 4.3) of the proposed architecture. In the following experiments, we focus on a very restricted proprioception-only setting implemented in the MIMo robotic-infant simulator (Mattern et al., 2024; López, Lenz, et al., 2025; López, Marcel, et al., 2025). For simplicity, the agent controls only one selected left-finger action variable, and the evaluated state variable is the corresponding proprioceptive position. Vision, vestibular sensing, tactile perception, and other modalities of the agent are disabled in these experiments, and the visual perceptual system developed in Section 4.2 is not evaluated here.

The purpose of these simple experiments is to evaluate three basic capabilities: predicting proprioceptive change from motor-babbling data (Section 5.2.2), inferring actions for desired proprioceptive targets (Section 5.2.3), and using internally sampled exploration targets to guide action inference (Section 5.2.4). The experiments are offline, i.e., the model is trained and evaluated on previously collected motor-babbling transitions, and the reported target-reaching results are computed through the learned

Table 5.6: Schema of the implemented real-time scene descriptor, listing the serialized keys exposed by the descriptor. Optional nested fields such as `bbox`, `pose`, and `velocity` are returned as $\emptyset$ when the corresponding evidence is unavailable.

Level	Keys	Value type
scene	`timestamp`, `frame_count`	scalar
scene	`num_objects`, `num_visible_objects`	integer
`objects[*]`	`obj_id`, `visible`, `age_s`, `last_updated`, `is_busy`	scalar / Boolean
`objects[*]`	`num_snapshots`, `num_poses`, `num_valid_poses`, `has_pose`	integer / Boolean
`bbox`	`xyxy`, `xywh`, `centroid`, `area`	integer vectors / scalar
`pose`	`position`, `orientation`, `timestamp`	float vectors / scalar
`velocity`	`linear`, `dt_s`	float vector / scalar
`objects[*]`	`last_pose_diagnostics`	JSON-compatible dictionary

WM rather than through repeated online simulator execution.

For the concrete implementation of this part, see Appendix A.2.

5.2.1 Experimental Configuration

The reported evaluation uses random motor babbling in the BabyBench/MIMo crib scene.[18] The simulator's built-in behavior controller is disabled, actuation uses the spring-damper model, and the simulator is stepped with a frame skip of 50. The motor configuration permits arm, hand, and finger actuation, but the analysis below is restricted to one scalar control variable and its corresponding scalar proprioceptive position.

In the notation of Section 4.3, the evaluated subsystem of $\mathcal{M}^{\mathrm{wm}}(t)$ treats the selected control variable as one component $a_{\mathrm{LF}}(t)$ of the action vector $\boldsymbol{a}(t)$, and the selected proprioceptive position as one component $q_{\mathrm{LF}}(t)$ of the proprioceptive state $\boldsymbol{q}(t)$. The motor-babbling dataset is extracted from the structured sensorimotor history $\mathcal{H}_t$ as a set of transition records of the form

$$\big[q_{\mathrm{LF}}(t-1), a_{\mathrm{LF}}(t), q_{\mathrm{LF}}(t)\big], \tag{5.1}$$

with $\Delta q_{\mathrm{LF}}(t) = q_{\mathrm{LF}}(t) - q_{\mathrm{LF}}(t-1)$ denoting the observed proprioceptive displacement.

The predictor used in this setup is a structured two-stage WM predictor rather than a single direct mapping from action to next state (akin to the notion of an FM; see Section 1.1.2 for comparison). In this proof-of-concept evaluation, the graph $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}$ is fixed manually from the selected transition relation rather than learned from data (i.e., $\mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}$ is expert-designed in the sense of Section 4.3.2). The resulting graph over the selected variables is shown in Fig. 5.5. Accordingly, the first learned local predictor estimates the displacement $\Delta q_{\mathrm{LF}}(t)$ from the action $a_{\mathrm{LF}}(t)$, and the second

[18] https://babybench.github.io/2025/api/#crib

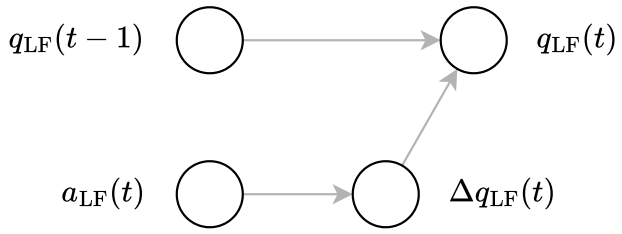


Figure 5.5: Graph of a simple proprioceptive WM operating only on one action variable $a_{\mathrm{LF}}(t)$. The action is mapped to proprioceptive displacement $\Delta q_{\mathrm{LF}}(t)$, which, together with the previous proprioceptive state $q_{\mathrm{LF}}(t-1)$, predicts the current proprioceptive state $q_{\mathrm{LF}}(t)$.

learned local predictor estimates the post-action position $q_{\mathrm{LF}}(t)$ from the pre-action position $q_{\mathrm{LF}}(t-1)$ and predicted displacement. Consequently, as per our formulation in Section 4.3.1,

$$\mathcal{V}^{\mathrm{act}} = \{a_{\mathrm{LF}}(t)\} \qquad \mathcal{V}^{\mathrm{obs}} = \{q_{\mathrm{LF}}(t-1)\} \qquad \mathcal{V}^{\mathrm{lat}} = \{\Delta q_{\mathrm{LF}}(t), q_{\mathrm{LF}}(t)\}. \tag{5.2}$$

In the reported implementation, the local mechanisms are additive Gaussian neural predictors, following the general additive-noise form (Eq. 4.31; Hoyer et al., 2008; Peters et al., 2017):

$$X_i \triangleq f_{\boldsymbol{\theta}_i}(\boldsymbol{x}_{\mathrm{pa}(X_i)}) + U_i, \tag{5.3}$$

but with $U_i \sim \mathcal{N}(0, \sigma_i^2)$; or, equivalently, the implied conditional distribution is

$$X_i \mid \boldsymbol{x}_{\mathrm{pa}(X_i)} \sim \mathcal{N}\left[f_{\boldsymbol{\theta}_i}(\boldsymbol{x}_{\mathrm{pa}(X_i)}), \sigma_i^2\right]. \tag{5.4}$$

Each local function $f_{\boldsymbol{\theta}_i}$ is implemented as a $\boldsymbol{\theta}_i$-parametrized multilayer perceptron (MLP; Rosenblatt, 1958, 1962; Minsky and Papert, 2017) with two hidden tanh-activated layers of 32 units, and a linear scalar output. The displacement predictor has one input, $a_{\mathrm{LF}}(t)$, whereas the post-action-position predictor has two inputs, $q_{\mathrm{LF}}(t-1)$ and $\Delta q_{\mathrm{LF}}(t)$.

This formulation instantiates the WM mechanism in a deliberately simple form. While the architecture allows local mechanisms to be parametrized by richer estimator families (e.g., by *normalizing flows*; Papamakarios et al., 2021; Winkler et al., 2019), the present proof-of-concept uses MLP regressors under an additive Gaussian residual model. The local predictors are trained separately on observed parent–child pairs, while evaluation uses the complete forward chain, so the post-action-position prediction includes any error propagated from the displacement prediction.

In total, the analysis uses 100,000 transitions from the dataset, split into an 80 : 20 train/test ratio. The observed range of the selected action variable is $a_{\mathrm{LF}} \in \mathcal{X}_a = [-1.000,\ 0.330]$,[19] and the observed range of the selected post-action proprioceptive

[19]Note that in our experimental configuration using the spring-damper muscle model, actions are dimensionless.

position is $q_{\mathrm{LF}} \in \mathcal{X}_q = [-2.754,\ 0.519]$ rad. The local predictors are trained with an MSE objective for 120 epochs with a batch size of 2,048 and a learning rate $\eta = 0.01$.

5.2.2 Experiment B1: Sensorimotor Transition Prediction

The first experiment evaluates whether the world model predicts the proprioceptive consequence of an action (i.e., whether it can perform a forward pass through the causal graph of $\mathcal{M}^{\mathrm{wm}}$). For each test transition, the model predicts both the action-induced displacement and the final post-action position:

$$\big[q_{\mathrm{LF}}(t-1), a_{\mathrm{LF}}(t)\big] \mapsto \big[\Delta\hat{q}_{\mathrm{LF}}(t), \hat{q}_{\mathrm{LF}}(t)\big]. \tag{5.5}$$

A persistence baseline is included as well, which predicts that the proprioceptive position remains unchanged, i.e., $\hat{q}_{\mathrm{LF}}(t) = q_{\mathrm{LF}}(t-1)$. This is a useful minimal comparison, since consecutive proprioceptive states are not independent: even without modeling the action, a predictor can obtain non-trivial accuracy by exploiting the temporal continuity of the joint state. Improving over this baseline, therefore, indicates that the learned WM captures action-dependent change rather than only reproducing the previous proprioceptive value.

Prediction error in this experiment is computed via MSE and MAE of $\hat{q}_{\mathrm{LF}}(t)$ and $\Delta\hat{q}_{\mathrm{LF}}(t)$ w.r.t. ground-truth values $q_{\mathrm{LF}}(t)$ and $\Delta q_{\mathrm{LF}}(t)$, respectively. During local training, the displacement-prediction objective $\mathrm{MSE}\,[\Delta\hat{q}_{\mathrm{LF}}(t),\ \Delta q_{\mathrm{LF}}(t)]$ decreased from the initial 0.653 rad^2 to 0.519 rad^2, while the post-action-position objective $\mathrm{MSE}\,[\hat{q}_{\mathrm{LF}}(t),\ q_{\mathrm{LF}}(t)]$ decreased from 0.380 rad^2 to 8.17×10^{-5} rad^2. Table 5.7 reports the test-set prediction errors for the individual prediction targets.

Table 5.7: Predictive performance on test motor-babbling transitions. Lower values are better.

Prediction target	MSE $[\mathrm{rad}^2]$	MAE $[\mathrm{rad}]$
Action-induced displacement	0.528	0.569
Post-action position	0.506	0.554
Persistence baseline	0.976	0.613

5.2.3 Experiment B2: Action Inference

The second experiment evaluates action inference for desired proprioceptive targets. The target set consists of 128 post-action proprioceptive values $q_{\mathrm{LF}}(t)$ from the test-set transitions. For each target, the corresponding pre-action state $q_{\mathrm{LF}}(t-1)$ is fixed. The target variable is the resulting proprioceptive value $X^{\mathrm{tar}} = q_{\mathrm{LF}}(t)$, corresponding to a selected variable in $\mathcal{V}^{\mathrm{tar}}(t)$.

Two action-selection conditions are compared. In the *random-action condition*, an action is sampled as

$$\tilde{a}_{\mathrm{LF}}(t) \sim \mathcal{U}(\mathcal{X}_a)\,, \tag{5.6}$$

where $\mathcal{X}_a$ denotes the observed action range defined in Section 5.2.1. The learned WM is used to predict the resulting post-action state from the same fixed pre-action state $q^*_{\mathrm{LF}}(t-1)$, i.e.,

$$\hat{q}_{\mathrm{LF}}(t) = \Phi_t\big[q^*_{\mathrm{LF}}(t-1), \tilde{a}_{\mathrm{LF}}(t)\big]. \tag{5.7}$$

This provides a control condition for target matching without inverse use of the WM.

In the *inferred-action condition*, the learned WM is kept fixed and the candidate action $\tilde{a}_{\mathrm{LF}}(t)$ is optimized directly via GD methods, following Algorithm 4.2. For each target, the optimization starts from the midpoint of the observed action range and minimizes the target-prediction error $\mathrm{MSE}\left[\hat{q}_{\mathrm{LF}}(t),\ q_{\mathrm{LF}}(t)\right]$ through the differentiable forward query Φ_t (Eq. 4.44). Candidate actions are updated by the Adam optimizer (Kingma & Ba, 2014) for 200 optimization steps with learning rate $\eta = 0.05$. After each update, the action is clipped to the observed action range $\mathcal{X}_a$; no additional action regularization is used.

The inferred actions produce lower target error than random actions. Across the 128 target states, the mean absolute target error decreases from 0.851 rad in the random-action condition to 0.195 rad in the inferred-action condition. The mean squared target-prediction objective decreases from 1.113 rad^2 at the beginning of action optimization to 0.286 rad^2 at the final step. In addition, 79.7% of inferred actions fall within a target error threshold of 0.05 rad, and 82.8% fall within 0.10 rad. Figure 5.6 visualizes the same comparison by plotting each target value against the post-action position predicted under the selected action. Points closer to the diagonal correspond to a more accurate target achievement.

5.2.4 Experiment B3: Curiosity-Conditioned Target Selection

The third experiment evaluates internally generated exploration targets. To select such targets, the observed range of the selected post-action proprioceptive variable $X = q_{\mathrm{LF}}(t)$ is discretized into 20 histogram bins. The training transitions are then used to estimate how often each bin has already been visited, yielding a simple empirical approximation of the outcome density $\hat{p}_{X,t}(x)$ as defined in Section 4.3.3. Before target

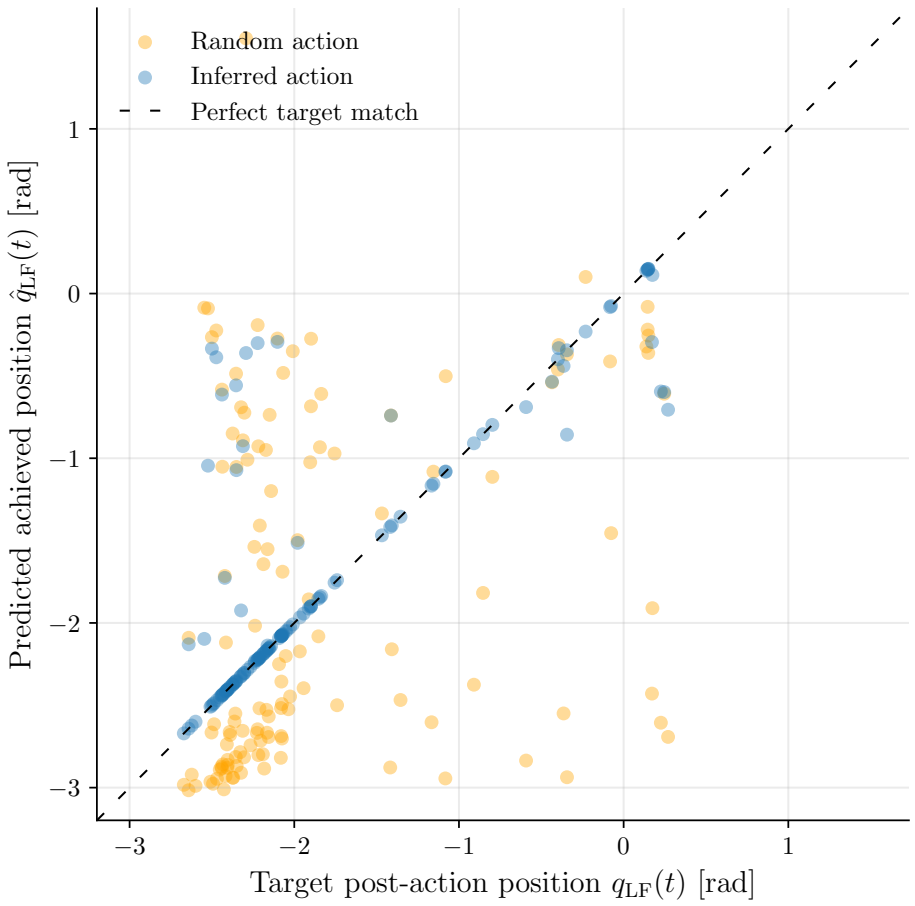


Figure 5.6: Performance of the action inference for proprioceptive targets. The *horizontal axis* shows the desired post-action position, and the *vertical axis* shows the post-action position predicted by the learned WM after applying either a random action or an inferred action. The *diagonal* indicates perfect target achievement.

generation, the training transitions occupy all 20 bins, with normalized entropy $H = 0.840$.

For each of 128 test-set pre-action states, the target sampler proposes one proprioceptive target from the corresponding curiosity-biased target density $p^{\mathrm{tar}}_{X,t}$. Let c_k denote the number of training observations in bin k, and let μ_k denote the bin center. The unnormalized probability of sampling a target from bin k follows then (as per Eqs. 4.34 and 4.36):

$$\tilde{p}^{\mathrm{tar}}_{k,t} = \left(\max_j c_j - c_k + \varepsilon\right)\left(1 + \alpha \exp\left[-\frac{1}{2}\left(\frac{\mu_k - q_{\mathrm{LF}}(t-1)}{\sigma}\right)^2\right]\right), \tag{5.8}$$

with

$$p^{\mathrm{tar}}_{k,t} = \frac{\tilde{p}^{\mathrm{tar}}_{k,t}}{\sum_j \tilde{p}^{\mathrm{tar}}_{j,t}}. \tag{5.9}$$

In the reported evaluation, $\alpha = 0.5$ and σ is set to $1/10$ of the observed position range. The inferred-action condition searches for one action per sampled target, whereas the

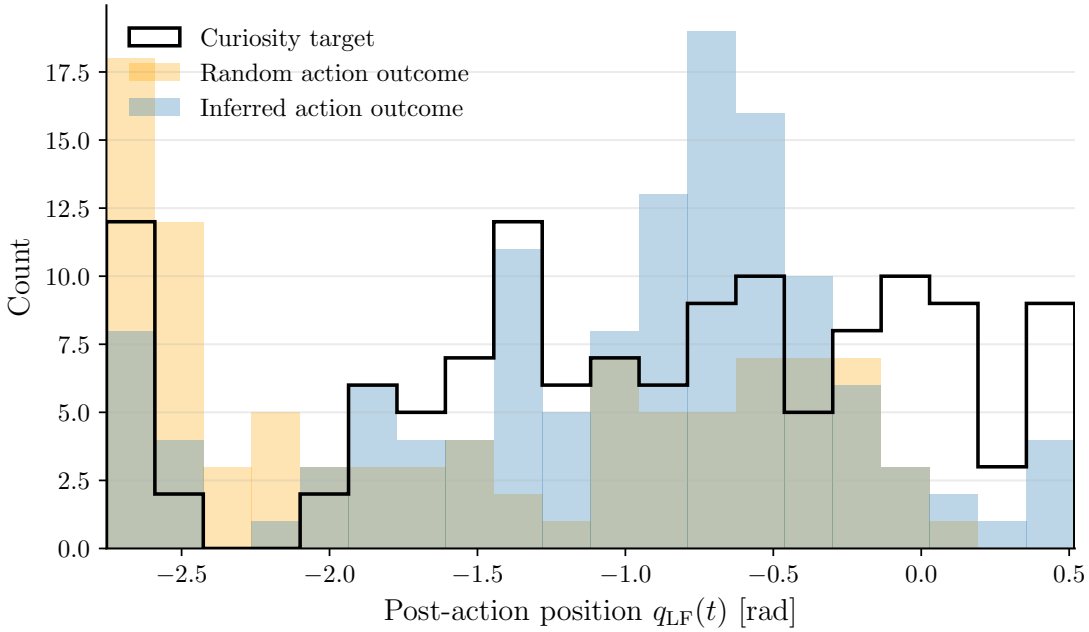


Figure 5.7: Curiosity-conditioned target selection and predicted proprioceptive outcomes. The *black outline* shows the sampled curiosity targets, while the *filled histograms* show the post-action positions predicted by the learned WM under random actions and under actions inferred for the sampled targets. All three histograms use the same 20-bin discretization of the selected post-action position $q_{\mathrm{LF}}(t)$.

random-action condition uses the same pre-action states and sampled targets, but replaces action inference with random action selection.

Under random action selection, the predicted outcomes occupy 18 of the 20 bins (occupancy ratio 0.90; normalized entropy $H = 0.866$). Under curiosity-conditioned action inference, the predicted outcomes occupy 19 of the 20 bins (occupancy ratio 0.95; normalized entropy $H = 0.900$). Target accuracy also improves: the target MAE decreases from 1.302 rad under random actions to 0.203 rad under curiosity-conditioned inferred actions. Figure 5.7 visualizes this comparison by showing the sampled curiosity targets together with the predicted outcomes produced by random actions and by inferred actions.

Chapter 6

Discussion

The experiments presented in Chapter 5 evaluate selected proof-of-concept components of the architecture specified in Chapter 4. Their purpose is not to validate the proposed architecture as a complete autonomous cognitive system, but to assess whether its central representational decomposition can be instantiated computationally. The results should therefore be interpreted primarily with respect to the aims defined in Section 3.1: transforming dynamic visual input into object-centric conceptual representations, learning structured sensorimotor action-effect regularities, and formulating both components within a common PAL (Varela et al., 1991; O'Regan & Noë, 2001; Noë, 2004; Fuster, 2004).

Overall, the experiments provide partial support for these aims. The perceptual subsystem (Section 5.1) demonstrates that RGB-D object observations can be converted into object records, retained snapshots, motion descriptors, visual descriptors, partial multidomain conceptual entities, and a structured scene descriptor. The WM subsystem (Section 5.2) then shows that a manually specified differentiable causal graph (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021) can be trained from proprioceptive motor-babbling transitions and queried for prediction, target-conditioned action inference, and curiosity-conditioned target pursuit. However, the two components were evaluated separately. Thus, the present thesis supports the feasibility of the proposed representational organization, but it does not yet empirically demonstrate a single integrated, online, multimodal agent system, in which the visual subsystem would supply structured information to the learned WM during closed-loop action.

6.1 Perception

The perceptual results shown in Section 5.1 to some degree demonstrate the viability of the visual subsystem as an object-centric representational interface between RGB-D input and downstream cognitive modeling. As specified in Section 4.2, the function of the perceptual subsystem is not to recover the complete state of the environment, but to estimate visually accessible state descriptors and organize them into structured object-level representations. The implemented system realizes this role only partially, but it nevertheless demonstrates an inspectable transformation from non-symbolic visual input to structured object records, trajectories, conceptual projections, and scene-level descriptors. This interpretation is also compatible with accounts of vision as an active, inferential process rather than a passive registration of the complete scene structure (Marr, 2010; Kersten et al., 2004; Yuille & Kersten, 2006; Goodale & Milner, 1992).

Exp. A1 (Section 5.1.2) evaluated the kinematic branch of this transformation. The results show that the current implementation can produce usable motion evidence in restricted conditions. Controlled translational motion of the textured box provided stable pose trajectories and could be embedded into the implemented motion space. This supports the architectural assumption that object motion can become a conceptual domain grounded in visual interaction, as suggested by active-perception and affordance-oriented accounts of object learning (Gibson, 2014; Fitzpatrick et al., 2003; Fitzpatrick, 2003). However, the same experiment also showed that this mechanism is not a robust solution to object motion perception. Because of the technical limitations of the kinematic pipeline, rotation and smooth-object motion revealed the current pose-estimation path's dependence on stable object-local features and reference snapshots. Consequently, the branch generally performs poorly at tracking objects without strongly discernible visual features, such as non-uniform texture or shape, because the feature-matching component underperforms in such cases, leading to its failure for pose estimation.

This limitation is nevertheless architecturally useful. The system distinguishes reliable geometric pose estimates from propagated or unavailable poses and records diagnostics for the pose-estimation process. As a result, unreliable motion evidence may be selectively excluded from the representation, while other features may still be retained to construct domain-specific meaning. This mechanism directly supports the partial-domain formulation presented in Section 4.2.2.6. Such selective information-domain availability is closer to the conceptual-spaces view of domain-dependent comparison (Gärdenfors, 2000, 2014; Bechberger & Kühnberger, 2017) than to an unconditional semantic vector representation in which all entities must be embedded into the same fixed feature set (Turney & Pantel, 2010; Mikolov et al., 2013).

Exp. A2 (Section 5.1.3) complements this result by showing that appearance-related conceptualization can remain available when motion evidence is absent. Qualitatively,

the ball record is the clearest example: it lacks a usable 6-DoF trajectory, but still has an object-local visual descriptor. This supports the separation between perceptual domains in the architecture. The retained EEMS snapshots function as object-local evidence, which, as mentioned in Section 4.2.3, is broadly compatible with object-file and visual-index accounts of temporary object-centered representation (Kahneman et al., 1992; Pylyshyn, 2001). At the same time, the visual descriptors should be interpreted as diagnostic appearance projections rather than as robust object-recognition embeddings. The prototype supports were very small, and the distances in Table 5.4 show that the descriptor does not provide clean object-label separation in all cases. For example, the box depth-translation record was closer to the single-exemplar ball prototype than to the box prototype. This is consistent with the descriptor's sensitivity to viewpoint, silhouette, visible surface structure, and segmentation quality. Thus, A2 supports the existence of an inspectable visual conceptualization mechanism, but not robust view-invariant identity recognition. Learned object-centric representation methods (Locatello et al., 2020; Stammer et al., 2024) could address part of this limitation in future work, although they would also reduce the immediate transparency of the handcrafted descriptor.

Exp. A3 (Section 5.1.4) provides the strongest evidence for the entity-level representational design. The final entity set consisted of five conceptual entities: two with both motion and visual projections, and three with only visual projections. This demonstrates the practical role of multidomain conceptual entities: objects can be represented across the domains for which evidence is available, whereas strict multidomain comparison is limited to entities that share the required domains. The comparison between the two box entities, with strict motion-visual distance $d = 9.31$, is therefore the only strict comparison over both implemented spaces. In terms of conceptual-space theory (Section 1.2.3), the system represents observed entities as exemplars in available domains and permits comparison only over relevant spaces (Medin & Schaffer, 1978; Rosch, 1978; Gärdenfors, 2000). This is a modest empirical result, given the small sample size and the prototypes' lack of robustness as category models, but it directly supports the intended representational structure.

The structured scene descriptor (Section 5.1.5) further supports the first aim of the thesis, albeit at an implementation-interface level rather than as a full scene-understanding result. The descriptor serializes the maintained object-centric state into fields such as object identifiers, visibility, bounding boxes, pose availability, velocity, snapshot counts, valid-pose counts, and pose diagnostics. Pairwise spatial relations, topological relations, and a scene grammar were not implemented, so this component should not be interpreted as the full scene-composition parser proposed in Section 4.2.2.5. Nevertheless, it demonstrates how the perceptual subsystem can expose its internal state in a form suitable for downstream modeling. The missing relational component is closely related to QSR and scene-graph approaches, in which object-

level descriptors are organized into explicit spatial, topological, and semantic relations (Cohn & Renz, 2008; Egenhofer & Franzosa, 1991; Randell et al., 1992; Johnson et al., 2015; Krishna et al., 2017; Chang et al., 2023).

In summary, the perceptual subsystem partially satisfies the first objective of Section 3.1. It constructs object-centric perceptual records, visual and motion conceptual projections, partial multidomain entities, and a structured scene descriptor. The main unresolved limitations are the absence of robust 6-DoF tracking under rotation and symmetry, the view sensitivity of the handcrafted visual descriptor, the lack of a relational scene grammar, and the absence of experimentally demonstrated object permanence or robust visual identity.

6.2 World Model

The WM (Section 5.2) experiments provide partial support for the second objective of Section 3.1. Their main contribution is the instantiation of a structured sensorimotor WM in which variables, graph topology, and local mechanisms are explicit. This follows the causal-modeling motivation that actions should be represented as intervention-like manipulations of modeled variables (Pearl, 2009; Peters et al., 2017; Hellström, 2021), while the mechanisms themselves may be learned predictors. The implemented model (Section 5.2.1) is, however, deliberately very minimal: it contains one scalar action variable, one scalar proprioceptive state variable, a manually specified graph, and additive Gaussian MLP mechanisms.

Exp. B1 (Section 5.2.2) evaluated forward prediction through the learned graph. The experiment indicates that the learned model captured some action-conditioned transition structure. Its improvement over a persistence baseline is important because proprioceptive states are temporally continuous; merely copying the previous state is already a non-trivial predictor. Beating this baseline suggests that the action variable contributes information about the next proprioceptive state. In this sense, the result is consistent with the role of forward models in sensorimotor prediction and body-schema learning (Wolpert & Kawato, 1998; Wolpert & Flanagan, 2001; Gama, 2026).

The prediction result is nevertheless mixed (Table 5.7). The displacement predictor remained relatively inaccurate, with test errors MSE = 0.528 rad^2 and MAE = 0.569 rad, and its training objective decreased only modestly. By contrast, the local post-action-position predictor reached a very low training error when trained with the observed displacement as input. This difference indicates that the difficult mechanism in the evaluated chain is the action-to-displacement mapping, not the additive update from the previous state and displacement to the post-action state. During full forward evaluation, errors in $\Delta\hat{q}_{\mathrm{LF}}(t)$ propagate into $\hat{q}_{\mathrm{LF}}(t)$, which makes the structured graph inspectable, enabling localizing where predictive failure enters the model, which is in line with the modularity assumption of structural causal models (Peters et al., 2017;

Vonk et al., 2023).

Exp. B2 (Section 5.2.3) evaluated target-conditioned action inference through the frozen trained WM. The experiment demonstrated that a differentiable WM under our definition (Section 4.3.1), while primarily acting as an FM, can be used inversely without any architectural changes, as we generated and optimized candidate actions in a backward pass through the learned graph so that their predicted consequences match a target. This outcome supports the action-generation mechanism proposed in Section 4.3.5 and connects to distal learning and model-based inverse control ideas, in which inverse action selection is solved using a forward predictive model rather than by learning a separate inverse mapping (Jordan & Rumelhart, 1992; Wolpert & Kawato, 1998; Baranes & Oudeyer, 2013). Nevertheless, the result remains model-internal. The optimized actions are evaluated using the learned WM, rather than through repeated online execution in an environment. Therefore, the experiment demonstrates the feasibility of a model-consistent action inference mechanism, not validated closed-loop control. This distinction is important because the WM bias can make inferred actions appear more accurate under the model than they would be under the true environment dynamics. A stronger result would require executing the inferred actions online and comparing realized proprioceptive outcomes with predicted ones.

Exp. B3 (Section 5.2.4) evaluated internally generated curiosity targets, essentially extending Exp. B2 from externally selected targets to internally sampled ones. The experiment was not intended to demonstrate body-schema exploration as described in Section 1.1.2 or by Gama (2026), as the one-dimensional target variable $q_{\mathrm{LF}}(t)$ was already well represented in the motor-babbling dataset; instead, the experiment demonstrates the computational coupling between intrinsic target generation and action inference (Sections 4.3.3 and 4.3.5, respectively). The target sampler proposes underrepresented or locally relevant proprioceptive outcomes, and the action-inference mechanism translates these targets into model-consistent candidate actions. This is related to intrinsically motivated exploration (Oudeyer & Kaplan, 2007; Schmidhuber, 2010; Schillaci et al., 2016; Aubret et al., 2023), but the present implementation does not yet realize a full active exploration loop. Such a loop would require online action execution in an interactive environment, observation of new action consequences, model updating, and subsequent revision of the curiosity distribution, closer to active BED and active causal-inference formulations as introduced in Section 1.3.3 and by Toth et al. (2022) and Gopnik and Schulz (2007).

Taken together, B1–B3 partially satisfy the second objective of Section 3.1. They demonstrate a minimal structured WM trained from motor-babbling transitions, forward prediction through a causal graph, inverse use of the FM for target-directed action inference, and curiosity-conditioned target pursuit. The claims remain bounded by the restricted setting, i.e., one scalar action, one scalar proprioceptive variable, an expert-designed graph, additive Gaussian local mechanisms, offline data usage, and

model-internal evaluation. Consequently, the results support the feasibility of the WM mechanism proposed in Section 4.3, but not full causal discovery, full active inference, or autonomous body-schema learning.

6.3 Perception-Action Integration

The empirical integration of the two implemented subsystems, as suggested in Section 4.1 and in the third objective of Section 3.1, remains incomplete. Nonetheless, the results support the perception and WM modules' compatibility at the level of representation. Both subsystems transform non-symbolic, continuous sensorimotor data into explicit, structured intermediate variables: the perceptual subsystem produces object-centric descriptors and conceptual projections, whereas the WM uses a graph structure to represent causal variables and relationships, and intervention-like action assignments, while the predictive local mechanisms embedded with the graph are implemented as MLPs. This shared structure is the main neurosymbolic connection between the two implementation blocks, consistent with neurosymbolic accounts that combine learned representations with explicit relational organization, as presented in Section 1.3.4 and by Garcez et al. (2019), Besold et al. (2021), and Butz et al. (2025).

Formally, the architecture assumes that the WM operates on a structured state estimate

$$\hat{\boldsymbol{s}}(t) = \hat{\boldsymbol{s}}^{\text{vis}}(t) \cup \boldsymbol{q}(t), \tag{6.1}$$

where $\hat{\boldsymbol{s}}^{\text{vis}}(t)$ is supplied by the perceptual subsystem and $\boldsymbol{q}(t)$ is directly observed proprioception. The perceptual experiments instantiate possible (but implementation-dependent) components of $\hat{\boldsymbol{s}}^{\text{vis}}(t)$, and the WM experiments instantiate the mechanism by which structured variables can support state and action inference. Thus, the two subsystems are compatible with the overall formalization in Section 4.1, even though they were not evaluated as one running system.

A stronger validation would require a closed perception-action experiment within an environment, in which object-level perceptual variables are included in $\mathcal{V}^{\text{obs}}$ or $\mathcal{V}^{\text{lat}}(t)$. The agent would then learn mechanisms relating its actions to visual, tactile, proprioceptive, or conceptual outcomes, and would select actions to achieve targets over these represented variables. Such an experiment would directly test whether the perceptual mapping $\Pi_{\boldsymbol{\theta}^{\text{perc}}}$ can provide evidence for $\mathcal{M}^{\text{wm}}(t)$ during embodied interaction. This remains the central point of future work.

6.4 Extensions

The following extensions are motivated by the limitations of the implemented proof of concept. They should therefore be understood as *hypothetical* continuations of the proposed architecture, not as empirical claims established by the present implementation.

6.4.1 World Model Hierarchization

The current WM experimental suite is restricted to a single proprioceptive action-effect relation. A full embodied agent would require variables spanning many joints, effectors, sensory modalities, objects, and timescales. A flat graph over all such variables would be difficult to learn, query, and interpret, as every prediction or action-inference problem would have to be represented at the same level of granularity. *Hierarchization* (Sridharan et al., 2019; Butz et al., 2025) could address this issue by organizing $\mathcal{M}^{\mathrm{wm}}$ into coupled submodels operating at different abstraction levels, analogously to hierarchical accounts of action control and action-sequence learning (Balleine et al., 2015; Eckstein & Collins, 2021).

A concrete implementation could, for instance, decompose the WM into a *lower sensorimotor layer* (LSML), an *intermediate body-event layer* (IBEL), and a *higher object-event layer* (HOEL). The LSML could contain predictive mechanisms over actuator commands, joint states, displacements, and local tactile variables. Variables of the LSML would then remain close to the form used in the present experiments, e.g., $a_i(t) \to \Delta q_i(t) \to q_i(t)$. The IBEL could summarize coordinated patterns over several such mechanisms, such as hand opening, reaching, self-touch, end-effector displacement, or contact onset. Formally, this can be understood as introducing *abstraction variables*

$$Z_k^{(l+1)}(t) = h_k^{(l)}\left[\boldsymbol{x}_{S_k}^{(l)}(t : t + \Delta t)\right], \tag{6.2}$$

where $S_k \subseteq \mathcal{V}^{(l)}$ denotes the support set, i.e., the subset of lower-level variables used to construct abstraction k, and $\boldsymbol{x}_{S_k}^{(l)}(t : t + \Delta t)$ denotes their values over a short temporal window. The function $h_k^{(l)}$ is the corresponding *abstraction mechanism*: it may be a learned temporal encoder, a skill or option detector, a dynamic motor primitive, or a rule-like event predicate mapping lower-level trajectories to a higher-level event, posture, or skill variable (Sutton et al., 1999; Ijspeert et al., 2013; Konidaris et al., 2018). The higher layer could then model relations among such events and object-level perceptual variables, for instance whether a reaching movement changes an object's position, produces contact, or brings an object into a reachable region. This would connect the proposed WM to hierarchical task-and-motion planning in robotics, where high-level symbolic goals are linked to lower-level continuous motion constraints (Kaelbling & Lozano-Pérez, 2011).

The flow of information could be bidirectional. Bottom-up, lower-level mechanisms would provide evidence for higher-level variables: repeated joint-level patterns could be compressed into motor primitives, contact events, or affordance-related outcomes. Top-down, higher-level targets would constrain lower-level action inference. For example, a target such as `touch(hand, object)` would not directly specify all joint commands. Instead, it would select an intermediate reaching/contact submodel[20], which would then generate lower-level proprioceptive or motor targets. Thus, action generation could be decomposed into nested queries, such as

$$X^{\text{tar}}_{\text{object}} \rightarrow Z^{\text{tar}}_{\text{event}} \rightarrow \boldsymbol{a}^*(t), \tag{6.3}$$

where object-level targets $X^{\text{tar}}_{\text{object}}$ (HOEL) are translated into event-level targets $Z^{\text{tar}}_{\text{event}}$ (IBEL) and finally into concrete actions $\boldsymbol{a}^*(t)$ (LSML). This mechanism is compatible with active-inference formulations in which higher levels encode more abstract, temporally extended expectations while lower levels specify faster sensorimotor predictions and actions (see Section 1.3.3 for more details; Friston, 2010; Friston et al., 2017; Parr and Friston, 2018; Parr et al., 2022). Similarly, the organization follows the developmental motivation discussed in Section 1.1, since early body-centered contingencies provide the basis for later object- and action-centered world modeling (Piaget, 1952; Gibson, 2014; Hellström, 2021; Gama, 2026).

Such hierarchization would also enable perceptual concepts to enter as higher-level observed or latent variables rather than as raw descriptor vectors. For example, object identity, approximate object position, motion class, contact relation, or affordance category could be represented as variables grounded in the perceptual subsystem but used by the WM for prediction and action inference. A low-level visual descriptor would hence not have to be a direct parent of every action-effect mechanism. Instead, it could support an HOEL variable such as `movable(object)` or `near(hand, object)`, which would then condition the relevant IBEL or LSML subgraph. This would preserve the link between continuous perception and structured inference while avoiding a monolithic transition model over raw sensory input, in accordance with grounded and multilevel accounts of semantic cognition (Barsalou, 1999; Lambon Ralph et al., 2016; Lake et al., 2016) and object-centric embodied learning in robotics (Zheng et al., 2025).

[20]This process could be assisted by a WM contextualization mechanism, such as the one proposed in Section 6.4.2.

6.4.2 World Model Contextualization

As the WM grows, inference should not necessarily involve the complete graph. A *context-specific query* may require only a subset of variables and mechanisms. *Contextualization* Butz et al., 2025 can therefore be formulated as the extraction of an active subgraph

$$\mathcal{G}^{C_t}_{\mathcal{M}^{\mathrm{wm}}}(t) \subseteq \mathcal{G}_{\mathcal{M}^{\mathrm{wm}}}(t), \tag{6.4}$$

where C_t denotes the current context, including the current state estimate, selected target variables, action constraints, uncertainty estimates, and possibly task or curiosity conditions. Importantly $\mathcal{G}^{C_t}_{\mathcal{M}^{\mathrm{wm}}}$ should be selected according to *causal relevance* (i.e., the relevance to the directed mechanisms that connect currently available variables to the queried targets) rather than by undirected statistical association alone.

For an additive-noise differentiable local mechanism formulation $X_i = f_i\left[\boldsymbol{x}_{\mathrm{pa}(X_i)}\right] + U_i$ as introduced in Section 4.3.1 and implemented in Section 5.2.1, the context-dependent relevance of a parent variable $X_j \in \mathrm{pa}(X_i)$ can be approximated, for example, by the *local sensitivity*

$$r^{C_t}_{j\to i} = \left| \frac{\partial f_i[\boldsymbol{x}_{\mathrm{pa}(X_i)}]}{\partial x_j}\bigg|_{\boldsymbol{x}_{\mathrm{pa}(X_i)}=\boldsymbol{x}^{C_t}_{\mathrm{pa}(X_i)}} \right|. \tag{6.5}$$

where $\boldsymbol{x}^{C_t}_{\mathrm{pa}(X_i)}$ denotes the parent assignment induced by the current context. Thus, $r^{C_t}_{j\to i}$ estimates how strongly the predicted value of X_i changes locally under an infinitesimal change of the parent X_j, while the remaining parent variables are held fixed. For vector-valued parents or children, the scalar derivative can be replaced by an appropriate Jacobian norm. This formulation is related to derivative-based sensitivity analysis (Sobol' & Kucherenko, 2009), while the restriction to directed parent–child mechanisms preserves the causal interpretation supplied by the structural graph (Pearl, 2009; Peters et al., 2017; Janzing et al., 2013).

The local score should not be interpreted as a complete *causal-effect estimate*. First, it is evaluated only at the current context, and may therefore miss influences that become relevant in other regions of the state space. Second, it measures sensitivity of the learned mechanism, whose accuracy depends on the available data and the chosen mechanism class. A more stable relevance estimate could average the local derivative over recent or sampled contexts, i.e.,

$$R^{C_t}_{j\to i} = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}\mid C_t)}\left[\left| \frac{\partial f_i[\boldsymbol{x}_{\mathrm{pa}(X_i)}]}{\partial x_j} \right| \right], \tag{6.6}$$

where $p(\boldsymbol{x} \mid C_t)$ denotes the agent's context-conditioned belief or empirical sampling distribution over WM states compatible with context C_t. Alternatively, we could replace the derivative with an ablation-based or interventional contrast when the mechanism is non-differentiable (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021). For

action inference, it may also be preferable to score relevance to the composed target query rather than to a single edge:

$$r^{C_t}_{j\to\mathrm{tar}} = \left\| \frac{\partial \Phi^{\mathrm{tar}}_t}{\partial x_j} \right\|, \tag{6.7}$$

where Φ^{tar}_t denotes the forward query (Eq. 4.44) from the current assignment to the selected target variables. This quantity estimates whether changes in X_j can affect the predicted target values through the currently active directed paths.

A contextualized query would then retain variables and mechanisms whose direct or path-level relevance exceeds a selected threshold, together with variables required to preserve directed paths from actions and observations to the target variables. Computationally, this reduces the cost of prediction and action inference in large graphs. Cognitively, it corresponds to activating only those sensorimotor variables that are relevant under the current situation. Moreover, contextualization could refine intrinsic motivation: curiosity targets may be sampled over uncertain or sparsely explored variables within the active subgraph rather than over all represented variables. *Empowerment*-like criteria (Gopnik, 2024) could also be incorporated at this level by preferring states from which the agent has many controllable future outcomes, connecting density-based curiosity to broader accounts of active developmental experimentation (Toth et al., 2022) and causal learning (Gopnik & Schulz, 2004; Gopnik & Schulz, 2007).

6.4.3 Artificial Formal Language of Thought

The explicit variables and mechanisms of the WM also motivate an artificial equivalent to the *language of thought* (LoT; Fodor, 1975; Rescorla, 2024) for internal reasoning. In this context, the LoT would not be a natural language, but a suitable alternative could be a *probabilistic-logic interface* for expressing queries, constraints, and relatively stable causal relations over grounded perceptual and sensorimotor variables. This preserves the classical idea that thought may use compositional internal representations (Fodor, 1975; Piantadosi et al., 2016), but replaces amodal atomic symbols with predicates grounded in perceptual descriptors, proprioceptive variables, and learned mechanisms. It is also compatible with probabilistic LoT accounts in which compositional representations denote probability distributions over possible world states (Goodman et al., 2015). Computationally, the same motivation appears in neurosymbolic systems, where learned components are combined with explicit relational or logical structure (Garcez et al., 2019; Besold et al., 2021; Lake et al., 2016; Bisták, 2025; Bisták & Farkaš, 2026).

The language could be defined over a typed signature

$$\Sigma_t \triangleq \langle \mathcal{O}_t, \mathcal{P}_t, \mathcal{A}_t, \mathcal{R}_t \rangle , \tag{6.8}$$

where $\mathcal{O}_t$ contains currently represented objects, agent's body parts, and relevant event tokens, $\mathcal{P}_t$ contains predicates such as `near`, `touch`, `reachable`, or `moving`, $\mathcal{A}_t$ contains action terms, and $\mathcal{R}_t$ contains typed relations among variables. A predicate would not be an uninterpreted symbol, but instead, each predicate $P \in \mathcal{P}_t$ could have a grounding map g_P, which

$$[\![P(o_1, \ldots, o_n)]\!]_t = g_P[\boldsymbol{x}_{S_P}(t), o_1, \ldots, o_n] \in [0, 1], \tag{6.9}$$

where $o_i \in \mathcal{O}_t$, S_P denotes the subset of perceptual or WM variables used to evaluate the predicate P (i.e., contextualization), and $\boldsymbol{x}_{S_P}(t)$ marks current values of that variable subset. For example, `near(hand, object)` could be grounded in relative spatial variables, whereas `touch(hand, object)` could combine spatial proximity, tactile evidence, and a contact-related WM variable. The truth value is therefore probabilistic (or graded), reflecting uncertainty in perception and prediction rather than a purely Boolean symbolic assertion. Such a design is close to probabilistic logic programming, where logical queries are evaluated under uncertainty (De Raedt et al., 2007; Manhaeve et al., 2021), and to statistical relational models such as Markov logic, where first-order formulas constrain probabilistic graphical structure (Richardson & Domingos, 2006; Koller & Friedman, 2009).

For instance, a *query* concerning a possible tactile contact outcome could be expressed abstractly as

$$\Pr[\texttt{touch}(\texttt{hand}, \texttt{object}) \mid \mathrm{do}[\texttt{move}(\texttt{hand}, \texttt{object})], \texttt{near}(\texttt{hand}, \texttt{object})]\,. \tag{6.10}$$

This expression should be understood as a query interface to the WM, not as a replacement for the learned mechanisms themselves. The term $\mathrm{do}[\texttt{move}(\texttt{hand}, \texttt{object})]$ denotes an intervention-like action assignment Pearl, 2009, while predicates in the conditioning context denote current grounded beliefs. Probabilistic logic would therefore provide a formal layer for composing uncertain perceptual facts with structural causal knowledge and action interventions operating within the WM (Pearl, 2009, 1988). In this setting, a rigid causal relation could be represented not as a weighted association such as `move` $\Rightarrow$ `touch`, but as a constraint tied to a directed mechanism in $\mathcal{M}^{\mathrm{wm}}$. For example, a *rule schema*

$$\texttt{reachable}(h, o) \wedge \mathrm{do}[\texttt{reach}(h, o)] \Rightarrow \Pr[\texttt{touch}(h, o)] > \kappa \tag{6.11}$$

would be licensed only if the corresponding subgraph contains mechanisms connecting the reach action, hand–object distance, and contact variable. Thus, the symbolic rule abbreviates a causal query over the WM.

Such a formal language could expose complex WM queries in an interpretable symbolic form, allow learned mechanisms to be combined with harder structural constraints, and provide a bridge between grounded conceptual representations and later linguistic or abstract reasoning. It would also make contextualization (Section 6.4.2)

more explicit, since a query such as `can_touch(hand, object)` could select the predicates, variables, and causal mechanisms required to evaluate that query, while ignoring unrelated parts of the graph. Its predicates would nevertheless have to remain grounded in the architecture's sensorimotor variables; otherwise, the language would reintroduce the *symbol-grounding problem* that the architecture is intended to partially address (Harnad, 1990; Vogt, 2002; Barsalou, 1999).

6.4.4 Natural Language Emergence

Natural language can be treated as a later communicative layer over grounded conceptual and causal representations (Roy, 2008). The present thesis does not implement language learning, symbol emergence, or language production. Nevertheless, the proposed architecture provides representational prerequisites for such mechanisms. Conceptual-space regions can ground labels for object properties and categories, while WM mechanisms can ground labels for actions, events, and causal relations (Tomasello, 2003; Bloom, 2000; Gärdenfors, 2024).

In this view, meaning is established prior to language through sensorimotor and perceptual organization. *Linguistic labels* could then function as pointers to regions, prototypes, or relations in the agent's grounded representational system (Section 4.2.2.6; Gärdenfors, 2024, 2000). For instance, regions of the conceptual visual domain may correspond to adjective-like expressions, kinematic-domain regions to verb-like expressions, and spatial or relational descriptors to preposition-like expressions. This could provide a possible route from grounded perception and action to compositional linguistic symbols, in accordance with the symbol-grounding motivation of Harnad (1990) and Vogt (2002).

This extension, however, remains substantially beyond the implemented system (Sections 5.1 and 5.2). A full model of language emergence would require mechanisms for assigning labels, coordinating them socially, resolving reference, composing expressions, and using linguistic input to modulate conceptual structure (Vygotskij, 1978; Lupyan, 2012). The present architecture would contribute only the pre-linguistic substrate, i.e., object-centric perceptual representations, conceptual spaces, and causal sensorimotor mechanisms to which later language could attach.

6.4.5 Sensorimotor Social Cognition

The object-centric and sensorimotor structure of the architecture also suggests a more speculative extension toward *social cognition*. Other agents may initially be represented as physical systems composed of body parts, poses, trajectories, and action effects, as per the theoretical capacities of the perceptual subsystem (Section 4.2). If it was extended to segment and track body parts, observed agents could be represented through kinematic and spatial descriptors analogous to those used for ordinary object tracking. This is consistent with developmental accounts in which causal and social understanding build on earlier action observation, imitation, and intervention-sensitive learning (Meltzoff et al., 2012; Waismeyer & Meltzoff, 2017; Gerstenberg & Tenenbaum, 2017).

The central additional requirement for this extension is a learned *self-model*. If the agent has acquired a body schema through proprioceptive and tactile exploration, it could compare observed body-part trajectories with its own controllable effectors (Rochat, 1998; DiMercurio et al., 2018; Gama, 2026). A simple self–other relationship could be formulated as a similarity comparison

$$\mathrm{dist}\left(\boldsymbol{c}^{\mathrm{obs}}_{\mathrm{limb}}, \boldsymbol{c}^{\mathrm{self}}_{\mathrm{effector}}\right), \tag{6.12}$$

where the two representations are embedded in a shared kinematic, spatial, or action-effect space. Such a mapping could support primitive *imitation* by identifying which of the agent's own action patterns is most similar to an observed movement. With a richer WM, it could also support elementary *intent inference* by querying which latent target or preference would make the observed action likely with respect to the agent's own WM.

Nevertheless, this extension should be interpreted cautiously. The present implementation contains no multi-agent setting, no body-part segmentation of other agents, no imitation-learning mechanism, and no theory-of-mind model. Moreover, social behavior cannot be reduced to deterministic physical motion, since other agents act according to internal goals, beliefs, and policies that are only partially observable. The relevance of the present architecture is therefore limited but structurally important: social cognition could be approached as an extension of embodied perception, self-modeling, and causal inference, rather than as an initially disembodied symbolic module.

Conclusion

This thesis proposed and partially implemented a minimal neurosymbolic perception-action architecture for embodied sensorimotor learning. The motivating assumption was that grounded cognitive representations should not be introduced as mature symbolic structures independent of perception and action. Instead, they should emerge from the organization of sensorimotor experience: the agent observes its environment, acts within it, receives new sensory consequences, and gradually constructs internal structures that support prediction, action generation, and further learning. The architecture developed in this work connects object-centric visual perception with a structured sensorimotor world model within a common perception-action loop.

The primary theoretical contribution of the work is the architectural formulation itself. The proposed system decomposes the embodied agent's processing into a perceptual subsystem and a world-model subsystem while keeping both components representationally compatible. The perceptual subsystem transforms raw RGB-D observations into structured object-level descriptors, retained object-local evidence, conceptual-space projections, and scene descriptors. The world model operates over structured variables and directed mechanisms, treats actions as intervention-like assignments, and supports prediction and target-conditioned action inference. This formulation combines continuous learned representations with explicit graph structure, causal variables, object records, conceptual domains, and interpretable queries. Therefore, it provides one possible route for connecting grounded perception, sensorimotor prediction, causal modeling, and later symbolic abstraction.

The computational contribution is a proof-of-concept evaluation of selected components of this architecture. The perceptual experiments demonstrated that recorded RGB-D object observations can be transformed into persistent object records, retained snapshots, visual and kinematic descriptors, partial multidomain conceptual entities, and structured scene descriptors. The results support the intended object-centric and multidomain organization, while also showing the limitations of the current visual subsystem, namely weak 6-DoF tracking under rotation, smooth-object motion, and poor feature support; view-sensitive handcrafted visual descriptors; and the absence of a relational scene grammar or a robust object permanence mechanism.

The world-model experiments demonstrated that a minimal differentiable causal model can be trained from proprioceptive motor-babbling data and queried in sev-

eral ways. The learned model captured an action-conditioned transition structure, although the difficult part was learning local mechanisms for non-trivial transformations. We showed that the same differentiable structure could then be used inversely for target-conditioned action inference, demonstrating that candidate actions can be optimized via the learned forward model. Finally, internally generated curiosity targets could be coupled to the action-inference mechanism, demonstrating the computational link between target sampling and model-consistent action generation. These results support the feasibility of the proposed world model mechanism under restricted conditions; however, they do not demonstrate full causal discovery, full active inference, autonomous body-scheme learning, or validated closed-loop control.

The main limitation of the thesis is that the two implemented components were not evaluated as one closed-loop multimodal agent embedded in an interactive environment. The perceptual subsystem produced object-centric descriptors, and the world model supported structured prediction and action inference, but visual variables were not inserted into the learned causal graph during online interaction. Consequently, the thesis supports the theoretical representational compatibility of the two subsystems, but not yet their empirical integration in a running perception-action system. A stronger validation would require an experiment in which actions change visually represented objects, the perceptual subsystem supplies object-level variables to the world model, and inferred actions are executed and evaluated against realized sensory consequences.

Several hypothetical extensions follow directly from this limitation. A larger agent suitable for more complex task and environmental settings would require a hierarchical world model, in which low-level sensorimotor variables would be abstracted into body events, object events, and affordance-related variables. Contextualization would be needed to select relevant world-model knowledge for a current query or target rather than evaluating the complete model. A formal artificial language of thought could provide a probabilistic-logic interface for expressing grounded queries and relatively stable causal constraints over the world model. Natural language and social cognition could then be treated as later layers atop grounded perceptual, conceptual, and causal representations rather than as initially disembodied modules.

In summary, the thesis shows that selected components of an embodied neurosymbolic perception-action architecture can be instantiated in an inspectable computational form. Its contribution is not a general robotic framework or a complete cognitive model, but a structured design and partial implementation demonstrating how object-centric perception, conceptual representation, causal world modeling, and action inference can be brought into a shared formal setting. The results, therefore, support the plausibility of the proposed representational decomposition and identify future work: the full integration of visual object representations and learned sensorimotor world models during active embodied interaction.

Bibliography

Adams, W. J., Graf, E. W., & Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Adolph, K. E., & Hoch, J. E. (2019). Motor development: Embodied, embedded, enculturated, and enabling. *Annual Review of Psychology*, *70*(70), 141–164. https://doi.org/10.1146/annurev-psych-010418-102836

Agrawal, R., Squires, C., Yang, K., Shanmugam, K., & Uhler, C. (2019). ABCD-Strategy: Budgeted experimental design for targeted causal structure discovery. In K. Chaudhuri & M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (pp. 3400–3409, Vol. 89). PMLR.

Aharon, N., Orfaig, R., & Bobrovsky, B.-Z. (2022). BoT-SORT: Robust associations multi-pedestrian tracking. https://doi.org/10.48550/arxiv.2206.14651

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, *C-23*(1), 90–93. https://doi.org/10.1109/t-c.1974.223784

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active vision. *International Journal of Computer Vision*, *1*(4), 333–356. https://doi.org/10.1007/bf00133571

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, *17*(8), 379–390. https://doi.org/10.1016/j.tics.2013.06.009

Andersen, R. A., & Buneo, C. A. (2002). Intentional maps in posterior parietal cortex. *Annual Review of Neuroscience*, *25*(1), 189–220. https://doi.org/10.1146/annurev.neuro.25.112701.142922

Anthwal, S., & Ganotra, D. (2019). An overview of optical flow-based approaches for motion segmentation. *The Imaging Science Journal*, *67*(5), 284–294. https://doi.org/10.1080/13682199.2019.1641316

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., & Yoshida, C. (2009). Cognitive developmental robotics: A survey. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 12–34. https://doi.org/10.1109/tamd.2009.2021702

Åström, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, *10*(1), 174–205. https://doi.org/10.1016/0022-247x(65)90154-x

Aubret, A., Matignon, L., & Hassas, S. (2023). An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, *25*(2), 327. https://doi.org/10.3390/e25020327

Augustyn, M., Frank, D. A., & Zuckerman, B. S. (2009). Infancy and toddler years. In W. B. Carey, A. C. Crocker, W. L. Coleman, E. R. Elias, & H. M. Feldman (Eds.), *Developmental-Behavioral Pediatrics* (4th ed., pp. 24–38). Elsevier. https://doi.org/10.1016/b978-1-4160-3370-7.00003-1

Ayzenberg, V., & Behrmann, M. (2024). Development of visual object recognition. *Nature Reviews Psychology*, *3*(2), 73–90. https://doi.org/10.1038/s44159-023-00266-w

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of multimodal large language models: A survey. https://doi.org/10.48550/arxiv.2404.18930

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94. https://doi.org/10.1111/j.0963-7214.2004.00281.x

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005. https://doi.org/10.1109/5.5968

Balleine, B. W., Dezfouli, A., Ito, M., & Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. *Current Opinion in Behavioral Sciences*, *5*, 1–7. https://doi.org/10.1016/j.cobeha.2015.06.001

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Baranes, A., & Oudeyer, P.-Y. (2013). Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, *61*(1), 49–73. https://doi.org/10.1016/j.robot.2012.05.008

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. https://doi.org/10.1017/s0140525x99002149

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

Bechberger, L., & Kühnberger, K.-U. (2017). Measuring relations between concepts in conceptual spaces. In M. Bramer & M. Petridis (Eds.), *Artificial Intelligence XXXIV* (pp. 87–100). Springer International Publishing. https://doi.org/10.1007/978-3-319-71078-5_7

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, *6*(5), 679–684. https://doi.org/10.1512/iumj.1957.6.56038

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/tpami.2013.50

Besold, T. R., d'Avila Garcez, A., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kühnberger, K.-U., Lamb, L. C., Lima, P. M. V., de Penning, L., Pinkas, G., Poon, H., & Zaverucha, G. (2021). Neural-symbolic learning and reasoning: A survey and interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press. https://doi.org/10.3233/faia210348

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, 3464–3468. https://doi.org/10.1109/icip.2016.7533003

Bisták, T. (2025). *Neuro-Symbolic Approach to Reinforcement Learning in Robotics* [Master's thesis, Comenius University Bratislava]. https://opac.crzp.sk/?fn=detailBiblioForm&sid=4013D2F8FE5692F45BC92E4F6B33

Bisták, T., & Farkaš, I. (2026). Neuro-symbolic reinforcement learning for robotics. https://doi.org/10.2139/ssrn.6537226

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Bloom, P. (2000). *How Children Learn the Meanings of Words*. The MIT Press. https://doi.org/10.7551/mitpress/3577.001.0001

Broadbent, D. E. (1977). Levels, hierarchies, and the locus of control. *Quarterly Journal of Experimental Psychology*, *29*(2), 181–201. https://doi.org/10.1080/14640747708400596

Busch, F. P., Willig, M., Zečević, M., Kersting, K., & Dhami, D. S. (2025). Structural causal circuits: Probabilistic circuits climbing all rungs of Pearl's Ladder of Causation. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=25XyUTICdZ

Butz, M. V., Mittenbühler, M., Schwöbel, S., Achimova, A., Gumbsch, C., Otte, S., & Kiebel, S. (2025). Contextualizing predictive minds. *Neuroscience & Biobehavioral Reviews*, *168*. https://doi.org/10.1016/j.neubiorev.2024.105948

Cambon, S., Alami, R., & Gravot, F. (2009). A hybrid approach to intricate motion, manipulation and task planning. *The International Journal of Robotics Research*, *28*(1), 104–126. https://doi.org/10.1177/0278364908097884

Cangelosi, A., & Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots* (L. B. Smith & R. C. Arkin, Eds.). The MIT Press.

Casiez, G., Roussel, N., & Vogel, D. (2012). 1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2527–2530. https://doi.org/10.1145/2207676.2208639

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*(3), 273–304. https://doi.org/10.1214/ss/1177009939

Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., & Hauptmann, A. (2023). A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 1–26. https://doi.org/10.1109/tpami.2021.3137605

Cibula, M. (2024). *Cognitively-Inspired Learning of Causal Relations in a Simulated Robotic Environment* [Bachelor's thesis]. Comenius University Bratislava. https://opac.crzp.sk/?fn=detailBiblioForm&sid=ACFB8EE4E98483452F48B506A168

Cibula, M. (2026a). *Causal Neurosymbolic Reasoning and Action Generation for Simulated Infants* (Version v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.20586486

Cibula, M. (2026b). *Constructing Grounded Conceptual Representations via Object-Centric Robotic Active Vision* (Version v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.20576012

Cibula, M., Kerzel, M., & Farkaš, I. (2024). Learning low-level causal relations using a simulated robotic arm. In M. Wand, K. Malinovská, J. Schmidhuber, & I. V. Tetko (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2024* (pp. 285–298). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72359-9_21

Cibula, M., Malinovská, K., & Kerzel, M. (2025). Towards bio-inspired robotic trajectory planning via self-supervised RNN. In W. Senn, M. Sanguineti, A. Saudargiene, I. V. Tetko, A. E. P. Villa, V. Jirsa, & Y. Bengio (Eds.), *Artificial Neural Networks and Machine Learning. ICANN 2025 International Workshops and Special Sessions* (pp. 149–160). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-04552-2_15

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/s0140525x12000477

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*(1), 1–36. https://doi.org/10.1080/095414400382181

Cohn, A. G., & Renz, J. (2008). Qualitative spatial representation and reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of Knowledge*

Representation (pp. 551–596, Vol. 3). Elsevier. https://doi.org/10.1016/s1574-6526(07)03013-1

Coninx, S. (2022). A multidimensional phenomenal space for pain: Structure, primitiveness, and utility. *Phenomenology and the Cognitive Sciences*, *21*(1), 223–243. https://doi.org/10.1007/s11097-021-09727-0

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*(2-3), 393–405. https://doi.org/10.1016/0004-3702(90)90060-d

Cowan, N. (1998). Attentional filtering and orienting. In *Attention and Memory: An Integrated Framework* (pp. 137–166). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195119107.003.0005

Culham, J. C., & Valyear, K. F. (2006). Human parietal cortex in action. *Current Opinion in Neurobiology*, *16*(2), 205–212. https://doi.org/10.1016/j.conb.2006.03.005

Dagum, P., & Luby, M. (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, *60*(1), 141–153. https://doi.org/10.1016/0004-3702(93)90036-b

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*, 886–893. https://doi.org/10.1109/cvpr.2005.177

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

De Raedt, L., Kimmig, A., & Toivonen, H. (2007). ProbLog: A probabilistic Prolog and its application in link discovery. *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 2468–2473.

Dearden, A. M., & Demiris, Y. (2005). Learning forward models for robots. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1440–1445.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences*, *8*(2), 66–70. https://doi.org/10.1016/j.tics.2003.12.004

Démuth, A. (2013). *Perception Theories* (A. Démuth, J. Dolista, S. Gáliková, P. Gärdenfors, R. Gray, M. Petrů, & A. Slavkovský, Eds.; 1st ed.). Faculty of Philosophy and Arts, Trnava University in Trnava; Towarzystwo Słowaków w Polsce.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*(1), 80–90. https://doi.org/10.1037/h0039515

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C.

Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

Dijkstra, N., Zeidman, P., Ondobaka, S., van Gerven, M. A. J., & Friston, K. (2017). Distinct top-down and bottom-up brain connectivity during visual perception and imagery. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-05888-8

DiMercurio, A., Connell, J. P., Clark, M., & Corbetta, D. (2018). A naturalistic observation of spontaneous touches to the body and environment in the first 2 months of life. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02613

Dodampegama, H., & Sridharan, M. (2023). Knowledge-based reasoning and learning under partial observability in ad hoc teamwork. *Theory and Practice of Logic Programming*, *23*(4), 696–714. https://doi.org/10.1017/s1471068423000091

Dogge, M., Custers, R., & Aarts, H. (2019). Moving forward: On the limits of motor-based forward models. *Trends in Cognitive Sciences*, *23*(9), 743–753. https://doi.org/10.1016/j.tics.2019.06.008

Douven, I., Verheyen, S., Elqayam, S., Gärdenfors, P., & Osta-Vélez, M. (2023). Similarity-based reasoning in conceptual spaces. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1234483

Eckstein, M. K., & Collins, A. G. E. (2021). How the mind creates structure: Hierarchical learning of action sequences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 618–624. https://escholarship.org/uc/item/4v65s77x

Edstedt, J., Bökman, G., Wadenbäck, M., & Felsberg, M. (2024). DeDoDe: Detect, Don't Describe – Describe, Don't Detect for local feature matching. *2024 International Conference on 3D Vision (3DV)*, 148–157. https://doi.org/10.1109/3dv62453.2024.00035

Edstedt, J., Bökman, G., & Zhao, Z. (2024). DeDoDe v2: Analyzing and improving the DeDoDe keypoint detector. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 4245–4253. https://doi.org/10.1109/cvprw63382.2024.00428

Egenhofer, M. J., & Franzosa, R. D. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems*, *5*(2), 161–174. https://doi.org/10.1080/02693799108927841

Eichenbaum, H. (2017). On the integration of space, time, and memory. *Neuron*, *95*(5), 1007–1018. https://doi.org/10.1016/j.neuron.2017.06.036

Elsayed, G., Mahendran, A., van Steenkiste, S., Greff, K., Mozer, M. C., & Kipf, T. (2022). SAVi++: Towards end-to-end object-centric learning from real-world videos. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh

(Eds.), *Advances in Neural Information Processing Systems* (pp. 28940–28954, Vol. 35). Curran Associates, Inc.

Engel, A. K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, *17*(5), 202–209. https://doi.org/10.1016/j.tics.2013.03.006

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, *12*(10), 388–396. https://doi.org/10.1016/j.tics.2008.07.004

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fagard, J., Esseily, R., Jacquey, L., O'Regan, K., & Somogyi, E. (2018). Fetal origin of sensorimotor behavior. *Frontiers in Neurorobotics*, *12*. https://doi.org/10.3389/fnbot.2018.00023

Farkaš, I., Malík, T., & Rebrová, K. (2012). Grounding the meanings in sensorimotor behavior using reinforcement learning. *Frontiers in Neurorobotics*, *6*. https://doi.org/10.3389/fnbot.2012.00001

Farkaš, I., Vavrečka, M., & Wermter, S. (2025). Will multimodal large language models ever achieve deep understanding of the world? *Frontiers in Systems Neuroscience*, *19*. https://doi.org/10.3389/fnsys.2025.1683133

Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., & Soatto, S. (2024). Multi-modal hallucination control by visual information grounding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14303–14312.

Ferraro, S., Mazzaglia, P., Verbelen, T., & Dhoedt, B. (2025). FOCUS: Object-centric world models for robotic manipulation. *Frontiers in Neurorobotics*, *19*. https://doi.org/10.3389/fnbot.2025.1585386

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Fitzpatrick, P. (2003). First contact: An active vision approach to segmentation. *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*, *3*, 2161–2166. https://doi.org/10.1109/iros.2003.1249191

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., & Sandini, G. (2003). Learning about objects through action - initial steps towards artificial cognition. *2003 IEEE International Conference on Robotics and Automation*, *3*, 3140–3145. https://doi.org/10.1109/robot.2003.1242073

Fodor, J. A. (1975). *The Language of Thought*. Crowell.

Francis, B. A., & Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, *12*(5), 457–465. https://doi.org/10.1016/0005-1098(76)90006-6

Frisby, S. L., Halai, A. D., Cox, C. R., Lambon Ralph, M. A., & Rogers, T. T. (2023). Decoding semantic representations in mind and brain. *Trends in Cognitive Sciences*, *27*(3), 258–281. https://doi.org/10.1016/j.tics.2022.12.006

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T. H. B., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/neco_a_00912

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T. H. B., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214. https://doi.org/10.1080/17588928.2015.1020053

Fuster, J. M. (2004). Upper processing stages of the perception–action cycle. *Trends in Cognitive Sciences*, *8*(4), 143–145. https://doi.org/10.1016/j.tics.2004.02.004

Gadre, S. Y., Ehsani, K., & Song, S. (2021). Act the part: Learning interaction strategies for articulated object part discovery. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15732–15741. https://doi.org/10.1109/iccv48922.2021.01546

Gallagher, S. (2017). *Enactivist Interventions: Rethinking the Mind*. Oxford University Press. https://doi.org/10.1093/oso/9780198794325.001.0001

Gama, F. (2026). *Learning Body Representations through Self-Touch: From Babies to Robots* [Doctoral dissertation, Czech Technical University in Prague].

Gama, F., Shcherban, M., Rolf, M., & Hoffmann, M. (2023). Goal-directed tactile exploration for body model learning through self-touch on a humanoid robot. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(2), 419–433. https://doi.org/10.1109/tcds.2021.3104881

Gao, M., Zheng, F., Yu, J. J. Q., Shan, C., Ding, G., & Han, J. (2023). Deep learning for video object segmentation: A review. *Artificial Intelligence Review*, *56*(1), 457–531. https://doi.org/10.1007/s10462-022-10176-7

Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. https://doi.org/10.48550/arxiv.1905.06088

Gärdenfors, P. (2000). *Conceptual Spaces: The Geometry of Thought*. The MIT Press. https://doi.org/10.7551/mitpress/2076.001.0001

Gärdenfors, P. (2006). Sensation, perception, and imagination. In *How Homo Became Sapiens: On the Evolution of Thinking* (pp. 25–53). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528517.003.0002

Gärdenfors, P. (2014). *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. The MIT Press. https://doi.org/10.7551/mitpress/9629.001.0001

Gärdenfors, P. (2019). Using event representations to generate robot semantics. *ACM Transactions on Human-Robot Interaction*, *8*(4), 1–21. https://doi.org/10.1145/3341167

Gärdenfors, P. (2020). Events and causal mappings modeled in conceptual spaces. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00630

Gärdenfors, P. (2024). Event structure, force dynamics and verb semantics. *Language Sciences*, *102*, 101610. https://doi.org/10.1016/j.langsci.2023.101610

Gärdenfors, P., & Lombard, M. (2018). Causal cognition, force dynamics and early hunting technologies. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00087

Gärdenfors, P., & Osta-Vélez, M. (2023). Reasoning with concepts: A unifying framework. *Minds and Machines*, *33*(3), 451–485. https://doi.org/10.1007/s11023-023-09640-2

Gärdenfors, P., & Williams, M.-A. (2001). Reasoning about categories in conceptual spaces. *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, 385–392.

Garrett, C. R., Chitnis, R., Holladay, R., Kim, B., Silver, T., Kaelbling, L. P., & Lozano-Pérez, T. (2021). Integrated task and motion planning. *Annual Review of Control, Robotics, and Autonomous Systems*, *4*(1), 265–293. https://doi.org/10.1146/annurev-control-091420-084139

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive Theories. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 515–548). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199399550.013.28

Gibson, J. J. (2014). The theory of affordances. In *The Ecological Approach to Visual Perception: Classic Edition* (1st ed., pp. 119–135). Psychology Press. https://doi.org/10.4324/9781315740218

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://www.deeplearningbook.org/

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts* (1st ed., pp. 623–654). The MIT Press. https://doi.org/10.7551/mitpress/9383.003.0035

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119. https://doi.org/10.1037/a0021336

Gopnik, A. (2024). Empowerment gain as causal learning, causal learning as empowerment gain: A bridge between Bayesian causal hypothesis testing and reinforcement learning. *Proceedings of the 29th Meeting of The Philosophy of Science Association*. https://philsci-archive.pitt.edu/id/eprint/24463

Gopnik, A., Meltzoff, A., & Kuhl, P. (1999). *The Scientist in the Crib: What Early Learning Tells Us about the Mind* (1st Perennial ed.). William Morrow and Company, Inc.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*(8), 371–377. https://doi.org/10.1016/j.tics.2004.06.005

Gopnik, A., & Schulz, L. E. (Eds.). (2007). *Causal Learning: Psychology, Philosophy, and Computation*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195176803.001.0001

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257–293). Cambridge University Press. https://doi.org/10.1017/cbo9780511752902.011

Gupta, A., & Davis, L. S. (2007). Objects in action: An approach for combining action understanding and object perception. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. https://doi.org/10.1109/cvpr.2007.383331

Ha, D., & Schmidhuber, J. (2018). Recurrent world models facilitate policy evolution. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, *31*, 2455–2467.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). Learning latent dynamics for planning from pixels. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (pp. 2555–2565, Vol. 97). PMLR. https://proceedings.mlr.press/v97/hafner19a.html

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. *Nature*, *640*(8059), 647–653. https://doi.org/10.1038/s41586-025-08744-2

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, *13*(79), 2409–2464.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, *56*(5), 872–876. https://doi.org/10.1037/h0040546

Hellström, T. (2021). The relevance of causation in robotics: A review, categorization, and analysis. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 238–255. https://doi.org/10.1515/pjbr-2021-0017

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*(5), 849–878. https://doi.org/10.1017/s0140525x01000103

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Howard, R. A. (1960). *Dynamic Programming and Markov Processes* (1st ed.). MIT Press.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems* (Vol. 21). Curran Associates, Inc.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, *8*(2), 179–187. https://doi.org/10.1109/tit.1962.1057692

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*. https://doi.org/10.48550/arxiv.2311.05232

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Ijspeert, A. J., Nakanishi, J., Hoffmann, H., Pastor, P., & Schaal, S. (2013). Dynamical movement primitives: Learning attractor models for motor behaviors. *Neural Computation*, *25*(2), 328–373. https://doi.org/10.1162/neco_a_00393

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., & Schölkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, *41*(5), 2324–2358. https://doi.org/10.1214/13-aos1145

Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11* (Version 11.0.0). https://github.com/ultralytics/ultralytics

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2015). Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668–3678. https://doi.org/10.1109/cvpr.2015.7298990

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307–354. https://doi.org/10.1207/s15516709cog1603_1

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1-2), 99–134. https://doi.org/10.1016/s0004-3702(98)00023-x

Kaelbling, L. P., & Lozano-Pérez, T. (2011). Hierarchical task and motion planning in the now. *2011 IEEE International Conference on Robotics and Automation*, 1470–1477. https://doi.org/10.1109/icra.2011.5980391

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219. https://doi.org/10.1016/0010-0285(92)90007-o

Kanazawa, H., Yamada, Y., Tanaka, K., Kawai, M., Niwa, F., Iwanaga, K., & Kuniyoshi, Y. (2022). Open-ended movements structure sensorimotor information in early human development. *Proceedings of the National Academy of Sciences*, *120*(1). https://doi.org/10.1073/pnas.2209953120

Katz, D., Kazemi, M., Andrew Bagnell, J., & Stentz, A. (2013). Interactive segmentation, tracking, and kinematic modeling of unknown 3D articulated objects. *2013 IEEE International Conference on Robotics and Automation*, 5003–5010. https://doi.org/10.1109/icra.2013.6631292

Kayhan, E., Heil, L., Kwisthout, J., van Rooij, I., Hunnius, S., & Bekkering, H. (2019). Young children integrate current observations, priors and agent information to predict others' actions. *PLOS ONE*, *14*(5), 1–16. https://doi.org/10.1371/journal.pone.0200976

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive Psychology*, *23*(2), 141–221. https://doi.org/10.1016/0010-0285(91)90009-d

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. https://doi.org/10.48550/arxiv.1412.6980

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026. https://doi.org/10.1109/iccv51070.2023.00371

Klein, G., & Crandall, B. W. (1995). The role of mental simulation in problem solving and decision making. In P. A. Hancock, J. M. Flach, J. Caird, & K. J. Vicente (Eds.), *Local Applications of the Ecological Approach to Human-Machine Systems* (1st ed., pp. 324–358). CRC Press. https://doi.org/10.1201/9780203748749-11

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques* (1st ed.). MIT Press.

Konidaris, G., Kaelbling, L. P., & Lozano-Perez, T. (2018). From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research*, *61*, 215–289. https://doi.org/10.1613/jair.5575

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32–73. https://doi.org/10.1007/s11263-016-0981-7

Kruzliak, A., Hartvich, J., Patni, S. P., Rustler, L., Behrens, J. K., Abu-Dakka, F. J., Mikolajczyk, K., Kyrki, V., & Hoffmann, M. (2024). Interactive learning of physical object properties through robot manipulation and database of object measurements. *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 7596–7603. https://doi.org/10.1109/iros58592.2024.10802249

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., & Sivic, J. (2022). MegaPose: 6D pose estimation of novel objects via render & compare. In K. Liu, D. Kulic, & J. Ichnowski (Eds.), *Proceedings of the 6th Conference on Robot Learning (CoRL)* (pp. 715–725, Vol. 205). PMLR.

Lake, B. M. (2014). *Towards More Human-Like Concept Learning in Machines: Compositionality, Causality, and Learning-to-Learn* [Doctoral dissertation, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/95856

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/s0140525x16001837

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2016). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Lauri, M., Hsu, D., & Pajarinen, J. (2023). Partially observable Markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, *39*(1), 21–40. https://doi.org/10.1109/tro.2022.3200138

Lee, T., Wen, B., Kang, M., Kang, G., Kweon, I. S., & Yoon, K.-J. (2025). Any6D: Model-free 6D pose estimation of novel objects. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11633–11643. https://doi.org/10.1109/cvpr52734.2025.01086

Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). EPnP: An accurate $O(n)$ solution to the PnP problem. *International Journal of Computer Vision*, *81*(2), 155–166. https://doi.org/10.1007/s11263-008-0152-6

Lin, J., Liu, L., Lu, D., & Jia, K. (2024). SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27906–27916. https://doi.org/10.1109/cvpr52733.2024.02636

Lindenberger, P., Sarlin, P.-E., & Pollefeys, M. (2023). LightGlue: Local feature matching at light speed. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17581–17592. https://doi.org/10.1109/iccv51070.2023.01616

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., & Gavves, E. (2023). BISCUIT: Causal representation learning from binary interactions. In R. J. Evans & I. Shpitser (Eds.), *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence* (pp. 1263–1273, Vol. 216). PMLR.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020). Object-centric learning with slot attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 11525–11538, Vol. 33). Curran Associates, Inc.

Lohmann, M., Salvador, J., Kembhavi, A., & Mottaghi, R. (2020). Learning about objects by learning to interact with them. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 3930–3941, Vol. 33). Curran Associates, Inc.

Lombard, M., & Gärdenfors, P. (2017). Tracking the evolution of causal cognition in humans. *Journal of Anthropological Sciences*, *95*, 219–234. https://doi.org/10.4436/JASS.95006

Long, Q., Scavino, M., Tempone, R., & Wang, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, *259*, 24–39. https://doi.org/10.1016/j.cma.2013.02.017

López, F. M., Lenz, M., Fedozzi, M. G., Aubret, A., & Triesch, J. (2025). MIMo grows! Simulating body and sensory development in a multimodal infant model. *2025 IEEE International Conference on Development and Learning (ICDL)*, 1–6. https://doi.org/10.1109/icdl63968.2025.11204381

López, F. M., Marcel, V., Hinaut, X., Triesch, J., & Hoffmann, M. (2025). BabyBench: A multimodal benchmark of infant behaviors for developmental AI.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Lupyan, G. (2012). What do words do? Toward a theory of language-augmented thought. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (pp. 255–297, Vol. 57). Academic Press. https://doi.org/10.1016/b978-0-12-394293-7.00007-8

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review*, *99*(4), 587–604. https://doi.org/10.1037/0033-295x.99.4.587

Mandler, J. M. (2007). *Foundations of Mind: Origins of Conceptual Thought*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195311839.001.0001

Manhaeve, R., Dumančić, S., Kimmig, A., Demeester, T., & De Raedt, L. (2021). Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence*, *298*, 103504. https://doi.org/10.1016/j.artint.2021.103504

Mao, W., Liu, M., & Salzmann, M. (2020). History repeats itself: Human motion prediction via motion attention. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 474–489). Springer International Publishing. https://doi.org/10.1007/978-3-030-58568-6_28

Mao, W., Liu, M., Salzmann, M., & Li, H. (2019). Learning trajectory dependencies for human motion prediction. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9488–9496. https://doi.org/10.1109/iccv.2019.00958

Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

Mattern, D., Schumacher, P., López, F. M., Raabe, M. C., Ernst, M. R., Aubret, A., & Triesch, J. (2024). MIMo: A multimodal infant model for studying cognitive development. *IEEE Transactions on Cognitive and Developmental Systems*, *16*(4), 1291–1301. https://doi.org/10.1109/tcds.2024.3350448

Maye, A., & Engel, A. K. (2012). Using sensorimotor contingencies for prediction and action planning. *From Animals to Animats 12*, 106–116. https://doi.org/10.1007/978-3-642-33093-3_11

McMains, S., & Kastner, S. (2011). Interactions of top-down and bottom-up mechanisms in human visual cortex. *The Journal of Neuroscience*, *31*(2), 587–597. https://doi.org/10.1523/jneurosci.3766-10.2011

Medin, D. L., & Coley, J. D. (1998). Concepts and categorization. In J. E. Hochberg (Ed.), *Perception and Cognition at Century's End* (pp. 403–439). Academic Press. https://doi.org/10.1016/b978-012301160-2/50015-0

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. https://doi.org/10.1037/0033-295x.85.3.207

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-old infants. *Developmental Psychology*, *48*(5), 1215–1228. https://doi.org/10.1037/a0027440

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*(8), 1265–1279. https://doi.org/10.1016/s0893-6080(96)00035-4

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Eds.), *1st International Conference on Learning Representations, ICLR 2013*. arXiv. https://doi.org/10.48550/arxiv.1301.3781

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, E. K., & Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science*, *263*(5146), 520–522. https://doi.org/10.1126/science.8290960

Minsky, M., & Papert, S. A. (2017). *Perceptrons: An Introduction to Computational Geometry* (S. A. Papert & L. Bottou, Eds.; 2017 ed.). The MIT Press. https://doi.org/10.7551/mitpress/11301.001.0001

Mirza, M. B., Adams, R. A., Mathys, C. D., & Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Frontiers in Computational Neuroscience*, *10*. https://doi.org/10.3389/fncom.2016.00056

Mota, T., Sridharan, M., & Leonardis, A. (2021). Integrated commonsense reasoning and deep learning for transparent decision making in robotics. *SN Computer Science*, *2*(4), 242. https://doi.org/10.1007/s42979-021-00573-0

Mumford, D. (1992). On the computational architecture of the neocortex: II The role of cortico-cortical loops. *Biological Cybernetics*, *66*(3), 241–251. https://doi.org/10.1007/bf00198477

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective* (4th print. ed.). MIT Press.

Myowa-Yamakoshi, M., & Takeshita, H. (2006). Do human fetuses anticipate self-oriented actions? A study by four-dimensional (4D) ultrasonography. *Infancy*, *10*(3), 289–301. https://doi.org/10.1207/s15327078in1003_5

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception*, *18*(1), 55–68. https://doi.org/10.1068/p180055

Neisser, U. (1967). *Cognitive Psychology*. Appleton-Century-Crofts.

Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. W. H. Freeman and Company.

Nematollahi, I., Mees, O., Hermann, L., & Burgard, W. (2020). Hindsight for foresight: Unsupervised structured dynamics models from physical interaction. *2020*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 5319–5326. https://doi.org/10.1109/iros45743.2020.9341491

Newell, A. (1994). *Unified Theories of Cognition* (3rd print., 1st Harvard Univ. Press paperback ed.). Harvard University Press.

Newell, F. N., McKenna, E., Seveso, M. A., Devine, I., Alahmad, F., Hirst, R. J., & O'Dowd, A. (2023). Multisensory perception constrains the formation of object categories: A review of evidence from sensory-driven and predictive processes on categorical decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1886), 20220342. https://doi.org/10.1098/rstb.2022.0342

Nguyen, P. D. H., Georgie, Y. K., Kayhan, E., Eppe, M., Hafner, V. V., & Wermter, S. (2021). Sensorimotor representation learning for an "active self" in robots: A model survey. *KI - Künstliche Intelligenz*, *35*(1), 9–35. https://doi.org/10.1007/s13218-021-00703-z

Nguyen-Tuong, D., & Peters, J. (2011). Model learning for robot control: A survey. *Cognitive Processing*, *12*(4), 319–340. https://doi.org/10.1007/s10339-011-0404-1

Noë, A. (2004). *Action in Perception* (1st ed.). MIT Press.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973. https://doi.org/10.1017/s0140525x01000115

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., & Contributors. (2024). *Computational Cognitive Neuroscience* (5th ed.). Online Book. https://book.compcogneuro.org

Osiurak, F., Rossetti, Y., & Badets, A. (2017). What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, *77*, 403–417. https://doi.org/10.1016/j.neubiorev.2017.04.014

Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*. https://doi.org/10.3389/neuro.12.006.2007

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, *22*(57), 1–64.

Parr, T., & Friston, K. J. (2018). The discrete and continuous brain: From decisions to movement – and back again. *Neural Computation*, *30*(9), 2319–2347. https://doi.org/10.1162/neco_a_01102

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. https://doi.org/10.1038/nrn2277

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the 7th Conference of the Cognitive Science Society*, 15–17.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (1st ed.). Morgan Kaufmann Publishers Inc. https://doi.org/10.1016/c2009-0-27609-4

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. https://doi.org/10.1145/3241036

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer* (1st ed.). Wiley.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (1st ed.). Basic Books.

Pech-Pacheco, J. L., Cristobal, G., Chamorro-Martinez, J., & Fernández-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: A comparative study. *Proceedings 15th International Conference on Pattern Recognition*, *3*, 314–317. https://doi.org/10.1109/icpr.2000.903548

Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, *3*(1), 13–26. https://doi.org/10.1038/s44159-023-00254-0

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North Ameri-*

can Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 2227–2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/n18-1202

Piaget, J. (1952). *The Origins of Intelligence in Children* (M. Cook, Ed.). International Universities Press. https://doi.org/10.1037/11494-000

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424. https://doi.org/10.1037/a0039980

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1), 127–158. https://doi.org/10.1016/s0010-0277(00)00156-6

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. In B. Nebel, C. Rich, & W. Swartout (Eds.), *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning* (pp. 165–176). Morgan Kaufmann Publishers Inc.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*. https://doi.org/10.48550/arxiv.2408.00714

Rescorla, M. (2024). The Language of Thought hypothesis. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2024). Metaphysics Research Lab, Stanford University.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*(1-2), 107–136. https://doi.org/10.1007/s10994-006-5833-1

Rochat, P. (1998). Self-perception and action in infancy. *Experimental Brain Research*, *123*(1-2), 102–109. https://doi.org/10.1007/s002210050550

Rolf, M., Steil, J. J., & Gienger, M. (2010). Goal babbling permits direct learning of inverse kinematics. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 216–229. https://doi.org/10.1109/tamd.2010.2062511

Rolls, E. T. (2016). Hierarchical organization. In *Cerebral Cortex* (pp. 40–71). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198784852.003.0002

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27–48). Lawrence Erlbaum. https://doi.org/10.4324/9781032633275-4

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

Roy, D. (2008). A mechanistic model of three facets of meaning. In *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 195–222). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199217274.003.0011

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., ... Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10105-3

Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*(1), 128–154. https://doi.org/10.1111/insr.12107

Saegusa, R., Metta, G., Sandini, G., & Sakka, S. (2009). Active motor babbling for sensorimotor learning. *2008 IEEE International Conference on Robotics and Biomimetics*, 794–799. https://doi.org/10.1109/robio.2009.4913101

Salge, C., Glackin, C., & Polani, D. (2014). Changing the environment based on empowerment as intrinsic motivation. *Entropy*, *16*(5), 2789–2819. https://doi.org/10.3390/e16052789

Schillaci, G., Hafner, V. V., & Lara, B. (2016). Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents. *Frontiers in Robotics and AI*, *3*. https://doi.org/10.3389/frobt.2016.00039

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247. https://doi.org/10.1109/tamd.2010.2056368

Schölkopf, B. (2022). Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (1st ed., pp. 765–804). Association for Computing Machinery. https://doi.org/10.1145/3501714.3501755

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634. https://doi.org/10.1109/jproc.2021.3058954

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332. https://doi.org/10.1111/j.1467-7687.2007.00587.x

Smallwood, R. D., & Sondik, E. J. (1973). The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, *21*(5), 1071–1088. https://doi.org/10.1287/opre.21.5.1071

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333. https://doi.org/10.1207/s15516709cog2803_1

Sobol', I., & Kucherenko, S. (2009). Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, *79*(10), 3009–3017. https://doi.org/10.1016/j.matcom.2009.01.023

Spaan, M. T. J. (2012). Partially observable Markov decision processes. In *Reinforcement Learning: State-of-the-Art* (pp. 387–414). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27645-3_12

Spalek, T. M., Unnikrishnan, K. P., & Di Lollo, V. (2024). Need for cross-level iterative re-entry in models of visual processing. *Psychonomic Bulletin & Review*, *31*(3), 979–984. https://doi.org/10.3758/s13423-023-02396-x

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. https://doi.org/10.1207/s15516709cog1401_3

Spelke, E. S., & Kinzler, K. D. (2006). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, *43*(6), 482–489. https://doi.org/10.1037/h0055479

Sridharan, M. (2020). REBA-KRL: Refinement-based architecture for knowledge representation, explainable reasoning and interactive learning in robotics. In G. De Giacomo, A. Catala, B. Dilkina, M. Milano, S. Barro, A. Bugarín, & J. Lang (Eds.), *Proceedings of the 24th European Conference on Artificial Intelligence* (pp. 2935–2936, Vol. 325). IOS Press. https://doi.org/10.3233/faia200461

Sridharan, M., Gelfond, M., Zhang, S., & Wyatt, J. (2019). REBA: A refinement-based architecture for knowledge representation and reasoning in robotics. *Journal of Artificial Intelligence Research*, *65*, 87–180. https://doi.org/10.1613/jair.1.11524

Srivastava, S., Fang, E., Riano, L., Chitnis, R., Russell, S., & Abbeel, P. (2014). Combined task and motion planning through an extensible planner-independent interface layer. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 639–646. https://doi.org/10.1109/icra.2014.6906922

Stammer, W., Wüst, A., Steinmann, D., & Kersting, K. (2024). Neural concept binder. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems* (pp. 71792–71830, Vol. 37). Curran Associates, Inc.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. https://doi.org/10.1038/nrn3838

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*(1-2), 181–211. https://doi.org/10.1016/s0004-3702(99)00052-1

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32. https://doi.org/10.1007/bf00130487

Takáč, M. (2007). *Construction of Meanings in Living and Artificial Agents* [Doctoral dissertation, Comenius University Bratislava]. https://dai.fmph.uniba.sk/~takac/thesis/

Teufel, C., Dakin, S. C., & Fletcher, P. C. (2018). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports*, *8*(1), 10853. https://doi.org/10.1038/s41598-018-28845-5

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, *28*(1), 287–313. https://doi.org/10.1023/a:1018912507879

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208. https://doi.org/10.1037/h0061626

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition* (1st Harvard Univ. Press paperback ed.). Harvard University Press. https://doi.org/10.2307/j.ctv26070v8

Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R., & von Kügelgen, J. (2022). Active Bayesian causal inference. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 16261–16275, Vol. 35). Curran Associates, Inc.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Umeyama, S. (1991). Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(4), 376–380. https://doi.org/10.1109/34.88573

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of Visual Behavior* (pp. 549–586). MIT Press.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.

Varela, F. J., Rosch, E., & Thompson, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press. https://doi.org/10.7551/mitpress/6730.001.0001

Vogt, P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, *3*(3), 429–457. https://doi.org/10.1016/s1389-0417(02)00051-7

von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip: Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Naturwissenschaften*, *37*(20), 464–476. https://doi.org/10.1007/bf00622503

von der Heydt, R. (2023). Visual cortical processing—From image to object representation. *Frontiers in Computer Science*, *5*. https://doi.org/10.3389/fcomp.2023.1136987

Vonk, M. C., Malekovic, N., Bäck, T., & Kononova, A. V. (2023). Disentangling causality: Assumptions in causal discovery and inference. *Artificial Intelligence Review*, *56*(9), 10613–10649. https://doi.org/10.1007/s10462-023-10411-9

Vygotskij, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Waismeyer, A., & Meltzoff, A. N. (2017). Learning to make things happen: Infants' observational learning of social and physical causal events. *Journal of Experimental Child Psychology*, *162*, 58–71. https://doi.org/10.1016/j.jecp.2017.04.018

Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., & Ding, G. (2025). YOLOE: Real-time seeing anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 24591–24602.

Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., & Birchfield, S. (2023). BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 606–617. https://doi.org/10.1109/cvpr52729.2023.00066

Wen, B., Yang, W., Kautz, J., & Birchfield, S. (2024). FoundationPose: Unified 6D pose estimation and tracking of novel objects. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17868–17879. https://doi.org/10.1109/cvpr52733.2024.01692

Whittington, J. C. R., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. J. (2020). The Tolman-Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263.e23. https://doi.org/10.1016/j.cell.2020.10.024

Winkler, C., Worrall, D., Hoogeboom, E., & Welling, M. (2019). Learning likelihoods with conditional normalizing flows. https://doi.org/10.48550/arxiv.1912.00042

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*(7–8), 1317–1329. https://doi.org/10.1016/s0893-6080(98)00066-5

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732. https://doi.org/10.1016/s0960-9822(01)00432-8

Woodward, J. (2011). A philosopher looks at tool use and causal understanding. In T. McCormack, C. Hoerl, & S. Butterfill (Eds.), *Tool Use and Causal Cognition* (pp. 18–50). Oxford University Press.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95. https://doi.org/10.1038/nature04262

Yao, R., Lin, G., Xia, S., Zhao, J., & Zhou, Y. (2020). Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology*, *11*(4), 1–47. https://doi.org/10.1145/3391743

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

Zappella, L., Lladó, X., & Salvi, J. (2008). Motion segmentation: A review. *Proceedings of the 2008 Conference on Artificial Intelligence Research and Development*. https://doi.org/10.3233/978-1-58603-925-7-398

Zareh, M., Toulabinejad, E., Manshaei, M. H., & Zahabi, S. J. (2024). A deep learning model of dorsal and ventral visual streams for DVSD. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-78304-7

Zhang, K., Schölkopf, B., Spirtes, P., & Glymour, C. (2017). Learning causality and causality-related learning: Some recent progress. *National Science Review*, *5*(1), 26–29. https://doi.org/10.1093/nsr/nwx137

Zheng, Y., Yao, L., Su, Y., Zhang, Y., Wang, Y., Zhao, S., Zhang, Y., & Chau, L.-P. (2025). A survey of embodied learning for object-centric robotic manipulation. *Machine Intelligence Research*, *22*(4), 588–626. https://doi.org/10.1007/s11633-025-1542-8

Zhou, T., Porikli, F., Crandall, D. J., Van Gool, L., & Wang, W. (2023). A survey on deep learning technique for video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7099–7122. https://doi.org/10.1109/tpami.2022.3225573

Zhu, S.-C., & Mumford, D. (2007). A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, *2*(4), 259–362. https://doi.org/10.1561/0600000018

Appendix A

Implementation Source Code

This thesis is accompanied by two code repositories containing the used implementations and experiments of the two major components of the proposed system:

- **Appendix A.1:** For the stable implementation release of the perceptual system (Section 4.2) and its experiments (Section 5.1), refer to Cibula (2026b) and to the associated code repository for the maintained version.
- **Appendix A.2:** For the stable implementation release of the world-model system (Section 4.3) and its experiments (Section 5.2), refer to Cibula (2026a) and to the associated code repository for the maintained version.