Scene Perception and Narrative Understanding by Multimodal Large Language Models

Branislav Zigo Comenius University Bratislava, Department of Applied Informatics branislav.zigo@fmph.uniba.sk

Abstract

Understanding an audiovisual perception as a story is a task that by far exceeds the complexity of image recognition speech-to-text conversion. and Such content understanding requires information that is not explicitly present in the scene, as well as estimation of the emotions, knowledge, and intentions of the actors involved. Decoding the meaning — understanding the story — can be considered as one of the most important human abilities. Humans have encoded this ability in their genome and develop it throughout life by observing, gaining experience, playing, learning, cooperating, and competing with others. (Mountford, Newbury 2019).

The degree of this ability largely determines a person's success and status in human society.

This article explores to which extent essential narrative structures, contextual cues, and commonsense knowledge can be made accessible to multimodal large language models to support coherent story comprehension.

1 Introduction

How do humans perceive a story? What cognitive faculties are necessary for narrative comprehension? Humans do not just passively absorb stories — they actively construct meaning through mental processes. During reading, hearing, or watching a story, human brain engages in (Bower, Morrow 1990):

A) Parsing the Structure - recognizing narrative elements: characters, setting, events, goals, conflicts, resolutions and understanding of the sequence and causality of events (e.g. what caused what).

B) Building a Mental Model - situation model or mental simulation. It is necessary for imagination of characters' goals, thoughts, emotions, and actions in a coherent world. ("What if?" question). This model updates dynamically as the story unfolds. **C) Inferring and Predicting** - filling in gaps in the narrative (inferences). Humans predict what might happen next based on prior knowledge and emotional cues.

D) Emotional and Empathic Engagement - empathy with characters. Emotional resonance helps encode and recall story content.

E) Integration with prior knowledge - stories are interpreted in context of the listener's experiences, culture, and expectations. A-priori created schemas help interpret ambiguous or symbolic content.

From the cognition point of view there are several prerequisites for story understanding. To comprehend a story, the brain needs (at least) the following cognitive abilities and conditions (Meylani 2024):

1. Language Comprehension - basic grammar, vocabulary, and syntax understanding. Even for visual stories (e.g. movies), narrative cues often rely on symbolic or linguistic elements.

2. **Theory of Mind** - ability to attribute mental states (beliefs, desires, emotions) to characters, or, more directly - ability to understand that other people have thoughts, beliefs, and emotions that may differ from one's own and from reality. It is critical for understanding motives, deception, irony, or complex plot twists.

3. **Causal Reasoning** - recognizing cause-and-effect relationships between events. This enables understanding of plot dynamics and tension.

4. **Temporal Sequencing** - understanding the order of events, even when the story uses flashbacks or non-linear storytelling.

5. **Working Memory** - holding elements of the story in mind while integrating new information. It is essential for understanding long or complex plots.

6. Schema Activation - access to relevant knowledge structures (schemas) — e.g., knowing

what a "birthday party" is helps you make sense of a story about one.

Story understanding is not passive — it's an active construction of meaning using memory, emotion, language, prediction, and social cognition.

It requires a complex (but mostly automatic) interplay between our linguistic, empathic, and reasoning systems.

When assessing Multimodal Large Language Models (MLLM) abilities to comprehend perceived scene and depicted story, we can inspect each cognition ability and look for strengths and weaknesses of the MLLMs. Some of the abilities (e.g. language comprehension, working memory, schema activation) can be considered as natural to MLLM due to its design. Some form of theory of mind is considered as existing in actual versions of GPT together with improvement of their language skills. Temporal sequencing and causal reasoning are currently subjects of research and development (Kosinski 2023).

2 Hypothesis and test methods

Multimodal large language models can be viewed as a distilled representation of human narrative knowledge, obtained primarily through the processing of literature. Literature deals with human stories, interpersonal relationships, and conflict resolution. With the MLLMs ability of image recognition, they should be able to comprehend narratives described in multimodal manner (Wang et. al. 2025).

The hypothesis: It can be expected that large MLLMs will be able to provide the context needed to understand events taking place in a scene and answer questions about the motivations and intentions of individual actors, they should be able to predict future development of the story and help to choose the optimal strategy for humans or an AI system if it is a participant in the story.

Several datasets with combined visual and language information can be used for evaluation and benchmarking of MLLMs abilities in scene recognition and narrative comprehension. As a sample can be mentioned:

a) MMIE - Massive Multimodal Interleaved Comprehension Benchmark (Xia et al. 2024)

- b) RAVEN Dataset for Relational and Analogical Visual Reasoning (Zhang et al. 2019)
- c) StripCipher dataset for testing the ability to comprehend and reason over sequences of images, evaluating temporal and contextual narrative understanding (Wang et al. 2025)
- d) CODIS Benchmark for Context-Dependent Visual Comprehension for Multimodal Large Language Models (Luo et al. 2024)
- e) VinaBench Benchmark evaluating Faithfulness and Consistency of generated Visual Narratives (Gao et al. 2025)

These datasets contain large amount of various test samples for statistical evaluation of MLLMs scene understanding and narrative comprehension.

Nevertheless, for the sake of conciseness and clarity, as well as for clear demonstration of strong and weak points of MLLMs in scene perception and narrative understanding, in this article we use just a few simple illustrative examples. Based on the results we will discuss root causes of MLLM success or failure in task fulfilment.

3 Simple experiments and used method

Experiments were performed with GPT-40 model (development status as of April 2025) and were focused on seeking weak points in model's decision making. One or more images together with text command or question were used. Images were intentionally labelled with meaningless labels. The aim was to check whether MLLM correctly understands the scene (at least for the purpose of the task), to force MLLM to find necessary context and assess whether model generates contextually appropriate responses. Each model answer was evaluated manually. The answer was checked for correctness, coherence, alignment with visual context and for intuitive similarity with human approach. In case of failure, we dive deeper, trying to find the root cause of the problem.

We used two group of tasks:

- 1. Content related tasks
- 2. Spatial orientation related task

The first group of tasks was related to the necessity to put content of the picture to the right context.

Task 1: Prompt: Imagine that you are barman in the pub. Person on the image asks you for cup of whisky with ice. Will you serve the person? Justify.



Fig. 1 Samples of pictures for "Barman" test (Freelicence pictures from unsplash.com)

The model works performs with high accuracy, applies legal restrictions and common-sense rules – refuses to provide alcohol to children, drunken persons, or even to visibly pregnant women. In case of doubt asks for age verification with ID. Can handle non-human pictures – e.g. primates.

It is readily to see that once MLLM properly understands person's attributes (e.g. age, physical status) on the picture, it is fully able to infer the right answer with proper justification.

Task 2: Prompt: What is happening on the picture? What would you advise to people nearby? Is it dangerous?



Fig. 2 Samples of pictures from "Fire hazard" test (Free-licence pictures from unsplash.com)

Again, the model works with high accuracy, correctly identifies nature of the fire as well as its emergency level. It provides proper advices based on picture assessment.

Next task allows ambiquous interpretations – usually there is no "solution". We try to challenge the model to interpret the scene and invent "solution". The purpose is to check whether offered answer is "human-like" or no.

Task 3 Human relationship. Prompt: *What is happening on the picture? How do you solve the situation?*

The model analysed each picture correctly and insightfully. It "proactively" offered plot creation based on the sequence of the pictures. Created plot was reasonable, understandable, however fairly predictable.



Fig.3 Violent scene (Free-licence picture from unsplash.com)

The model commented Fig.3 as follows: "The scene depicted appears to portray a very serious and distressing situation: a woman seems to be forcibly restrained by a man who has his hand over her mouth, possibly indicating an act of violence or coercion. Her wide, fearful eyes and his closed eyes and forceful grip suggest a non-consensual encounter or an abduction-like scenario. What to do in such a situation (in real life): If you or someone you witness is in immediate danger:

Call emergency services immediately (e.g., 911 or local equivalent).

Do not try to intervene physically unless it is safe to do so — instead, find help, document details, or alert others nearby.

If you are the person in danger, look for a chance to escape, create noise, or use an emergency feature on a phone if possible (many smartphones allow pressing the power button several times to call for help).

After the fact, seek medical attention and report the incident to authorities. Support services, such as

hotlines and shelters, can also help. Important:

If this image is part of a project or you're working on awareness, education, or safety training, I can assist further. Otherwise, if this is a real-life concern, please don't hesitate to seek immediate help."

Other samples are less violent and model properly estimates not only facial expressions, emotions of the actors, but also meaning of symbols (black clothes, red rose, weapons, etc...)



Fig. 4 Examples of human relationship (Free-licence pictures from unsplash.com)

There is a visible model behaviour pattern, that at first it recognizes persons, objects, emotions and symbols on the picture and in the second stage uses these extracted tokens for further inference and reasoning.

The second group of tasks was related to spatial dependencies of the objects on the picture. The purpose of the scenarios was to evaluate spatial awareness and to solve situation from physical world.



Fig. 5 Samples from driving licence tests (Tests from Slovak ministry of internal affairs <u>https://www.minv.sk/?testy</u>)

Task 4 Prompt: *"Write down the order in which cars pass the crossroad. Justify your answer."*

The results from these tests were total failure. MLLM performed extremely bad. Success rate was below 10%. The model properly recognized each object and has been aware of traffic rules. It tried to infer proper solution based on information extracted from the picture. Nevertheless it failed even on "basic" issue – inability to properly recognize left and right side of the car based on its orientation on the picture (It means that model evaluates left and right side as a feature of the whole picture, not the feature related to object on the picture – when the object turns, its left and right side turns also, but this cause spatial schema mismatch). It looks like a model spatial (un)awareness is fundamental gap in inference and reasoning.

To investigate this phenomenon further, the following trivial task was given:



Fig. 6 Simple colour strips

Task 5a Prompt: "Read the colours from left to right"

Task 5b Prompt: "Read the colours from right to left"

While task 5a was executed properly, task 5b was performed wrongly. Even such a seemingly simple task was out of abilities of the model. Asking for basic mutual objects relationships (left-right from, upper-lower than...) also led to suboptimal results and significant number of mistakes occurred:



Fig.8 Test of mutual object position awareness II

All objects on the pictures were recognized properly – both shape and colour, as well as their absolute position within the picture. But when asked which object is upper or lower than other object or more left or right than other object the number of mistakes was more than 20%. It is worth to notice, that significantly more mistakes were related to left-right relationship than upper-lower relationship.

To dig deeper we tried simple physics related questions:



Fig. 9 Weights - samples

Task 9a Prompt: *"Which picture is physically correct and why?"*

Task 9b Prompt: "*Put the picture 1 and picture 3 to the right order timewise. Justify your decision.*"

In Task 9a model selected wrong picture – with correct justification. Task 9b was solved properly with proper justification. The first picture was correctly described as in unstable state leading to the movement of the weights to stable state.

Similar examples with pulleys (question whether weights will move and in which direction) were solved properly without any mistake.



Fig. 10 Samples of pulleys with weights

But more complex task with atmospheric pressure was not solved correctly, even the model seemingly understood the concept of the air pressure and barometer and its justification has been done properly. The experiment has been designed with two mercury barometers. Mercury should compensate force created by atmospheric pressure by its own weight – the right picture is correct. It seems that the problem is again hidden in spatial unawareness – the model simply cannot distinguish correct mutual object relations from incorrect ones.



Fig. 11 Sample with barometers

4 Analysis of the results

Presented test results show significant imbalance in MLLM abilities. The model excels in extraction of the symbols from the pictures and their combinations with the text tokens. It seems that it can effectively simulate some of the key human abilities:

- Language Comprehension
- Theory of Mind
- Causal Reasoning
- Temporal Sequencing
- Working Memory
- Schema Activation

Unfortunately, it shows huge gaps in several basic spatial abilities:

- Spatial perception
- Spatial relation
- Spatial orientation
- Mental rotation
- Spatial visualization

These abilities can be summarized under more general term of spatial awareness. Lack of spatial awareness is most likely caused by non-existence of physical world related training data for GPT. It hinders its ability to understand narratives strongly related to 2D schemas and 3D physical world.

This gap is often expressed as inability to distinguish between left and right side of the object or in improperly perceived mutual position of the objects.

Conclusion and future work

MLLM models excel in visual object recognition and feature extraction. They can effectively combine this information with their textual knowledge base, which allows them in many cases to approximate human-like behaviour in scene recognition and narrative understanding. Therefore, we can see our hypothesis as partially confirmed.

Unfortunately, the lack of spatial awareness (missing grounding in the physical world) causes strong imbalance in their abilities (Xu et al. 2025)

These results are in line with results from the attempts to solve science and engineering problems with ChatGPT (Wang et al. 2024)

Due to this gap, current MLLMs are very often not able to solve even simple real-life situations. To mitigate this issue, large amount of training data related to spatial abilities will be necessary. However, the ability of model reasoning based on these data remains a question.

Acknowledgement

We thank for support to the Slovak Society for Cognitive Science. <u>https://cogsci.fmph.uniba.sk/sskv/</u>

References

- Wenrui Xu, Dalin Lyu, Weihang Wang, Jie Feng, Chen Gao, Yong Li (2025). Defining and Evaluating Visual Language Models' Basic Spatial Abilities: A Perspective from Psychometrics. https://doi.org/10.48550/arXiv.2502.11859
- Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, Huaxiu Yao (2024).
 MMIE: Massive multimodal interleaved comprehension benchmark for large Vision-Language Models. https://doi.org/10.48550/arXiv.2410.10139
- Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, Song-Chun Zhu (2019). RAVEN: A Dataset for Relational and Analogical Visual rEasoNing. https://doi.org/10.48550/arXiv.1903.02741
- Xiaochen Wang, Heming Xia, Jialin Song, Longyu Guan, Yixin Yang, Qingxiu Dong, Weiyao Luo, Yiru Wang, Yifan Pu, Xiangdi Meng, Wenjie Li, Zhifang Sui (2025). Beyond Single Frames: Can LMMs Comprehend Temporal and Contextual Narratives in Image Sequences? https://doi.org/10.48550/arXiv.2502.13925
- Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, Yang Liu (2024).
 CODIS: Benchmarking Context-Dependent Visual Comprehension for Multimodal Large Language Models. https://doi.org/10.48550/arXiv.2402.13607
- Silin Gao, Sheryl Mathew, Li Mi, Sepideh Mamooler, Mengjie Zhao, Hiromi Wakaki, Yuki Mitsufuji, Syrielle Montariol, Antoine Bosselut. (2025). VinaBench: Benchmark for Faithful and Consistent Visual Narratives. https://doi.org/10.48550/arXiv.2503.20871
- Thomas C. Foster (2014) How to read literature like a professor – a lively and entertaining guide to reading between the lines, revised edition. Harper Perennial. Slovak translation (2018)

David Foster (2023) Generative Deep Learning – Teaching Machines to Paint, Write, Compose and Play. Second Edition. O'Reilly

Michal Kosinski (2023) Theory of Mind May Have Spontaneously Emerged in Large Language Models. https://doi.org/10.48550/arXiv.2302.02083

Hayley S. Mountford, Dianne F. Newbury (2019) The Genetics of Language Acquisition.
<u>https://www.researchgate.net/publication/33305657</u>
<u>9 The Genetics of Language Acquisition</u>

Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2023). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. Frontiers in Education, 8, Article 1330486. <u>https://doi.org/10.3389/feduc.2023.1330486</u>

Gordon H. Bower, Daniel G. Morrow (1990) Mental Models in Narrative Comprehension <u>https://www.jstor.org/stable/2873581?origin=JSTO</u> <u>R-pdf</u>

Rusen Meylani (2024) Innovations with Schema Theory: Modern Implications for Learning, Memory, And Academic Achievement. https://doi.org/10.36948/ijfmr.2024.v06i01.1378