Schema Perception for Robust Video Question Answering

Xijun Wang[1,2], **Linlin Lu**[1], **Junbang Liang**[1], **Chun-Kai Wang**[1], **Kenan Deng**[1], **Chris Wang**[1], **Son Tran**[1], **Arnab Dhua**[1], **Michael (Yu) Lou**[1], **Ming Lin**[1,2], **Shan Yang**[1]
[1]Amazon, [2]University of Maryland, College Park
Palo Alto, CA 94301
{xijunw, ssyang}@amazon.com

Abstract

Recent advances in multi-modal perception that combines multi-modal input processing suffers from a common issue that the models predict something out of nothing or the results are inconsistent with the facts. In particular, in the context of visual question and answering. Many researches have done to reduce this problem by offering better vision-language alignment or visual prompting for visual related tasks. However, the previous methods still suffers from miss aligned attentions such as the model is paying attention to the wrong part of the visual input or miss certain part of the visual input, especially for question and answering task. Human brain, on the other hand is great at sourcing information and paying attention to the part of the visual input based on the given question. In this paper, we proposed a new framework that initiated from the human perception psychology, the Schema theory. Schema theory explains how human naturally organize and interpret external and internal information and incorporating them for reasoning and decision-making. We apply the Schema theory to dynamically prompt the multi-modal models to maximumly reduce the poor perception performance problem. From our experiments, ChatGPT-4o with SoT achieved 13.3% improvement, Claude 3.5 Sonnet with SoT achieved 28% improvement on MMVP. We found our Schema of Thought framework consistently improved the performance on public hallucination benchmarks.

1 Introduction

Recent advances in Multimodal Large Language Model(MLLMs) [7, 15, 32, 23, 2, 13] has revolutionalized the cross-modality learning. Especially the visual question and answering (VQA) problems. The language reasoning capability in MLLMs is one of the major reason behind this advances. As shown in Figure 1, however, many of the state-of-the-art MLLMs still suffers from the halllucination problem as discussed in many papers [25]. There are a couple of hallucination reduction prompting techniques, such as the language only chain-of-thought (COT) [27] prompting technique and the following up work such as the Tree-of-Thought (ToT) [31] and external world knowledge-based [18, 21]. There are also work that try to prevent multi-modal hallucination like prompting with visual cues [29], using scene understanding [22]. With the research in prompting engineering, the other challenge emerges, the efficiency of the prompting [33]. Human on the other hand is great at processing perceived information. Human being can efficiently leveraging knowledge in the memory to process and direct attention to the related visual information when prompted with a text question. This particular mechanism however hasn't been well studied for the cross-modality hallucination reduction problem.

In this paper, we propose a human cognitive psychology-based prompting techinque, Schema of Thought (SoT) targeting at reducing the hallucination problem of the MLLMs. In the Schema

12th Amazon Consumer Science Conference (CSS 2025), San Diego, CA. Amazon Confidential.

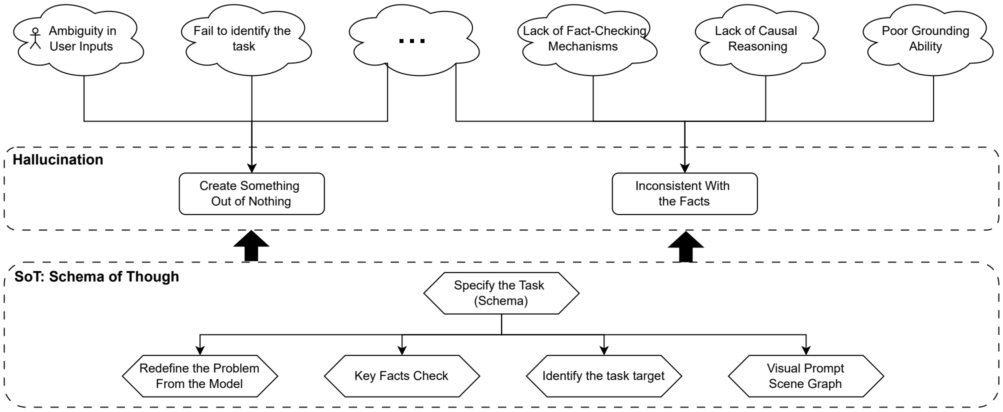
Ambiguity in User Inputs
Fail to identify the task
...
Lack of Fact-Checking Mechanisms
Lack of Causal Reasoning
Poor Grounding Ability
Hallucination
Create Something Out of Nothing
Inconsistent With the Facts
SoT: Schema of Though
Specify the Task (Schema)
Redefine the Problem From the Model
Key Facts Check
Identify the task target
Visual Prompt Scene Graph

Figure 1: **Task Overview.** Hallucination includes creating something out of nothing or inconsistent with the facts.

theory [3, 19], the current sensory input is a function of the subject's exploration of the world. In another word, human interpret new information by fitting it into existing frameworks or ”schemas“, which are organized patterns of thought and behavior. In a multi-modal information processing scenario, when a human is presented a text question and consequently an image, in Schema theory, human brain relies on the schema created from the text question to connect visual cues. This Schema mechanism also greatly reduces cognitive load by simplifying the information processing task. Instead of analyzing every visual detail independently, schemas allow the system to interpret complex scenes more efficiently by focusing on familiar patterns. As shown in Figure 2, we propose a Schema Theory based prompting technique that mimic how human leverage their knowledge in memory when processing new information. We see our framework SoT greatly reduces the hallucination on multiple benchmark datasets.

Our contribution can be summarized as, 1) Schema of Thought, a psychology-based framework for visual question and answering task; 2) we show through extensive experiment that our SoT framework is effective and robust for downstream visual QA task; 3) we improve consistently upon the Eyes Wide Shut (MMVP), HallusionBench and AutoHallusion datasets for real challenging test cases.

2 Related Works

2.1 Imitating Human Perception

Ever since human started to learn the biological and psychological self, there have been countless efforts on modeling how human perceive and reason about the world.

Gestalt Theory [28] states that human understands objects and concepts in top-down level, not the other way round. The Gestalt psychologists were the first to systematically study perceptual grouping [8], and found that it is easier for people to learn things that are regular, orderly, symmetrical, and simple. Although it was later criticized by modern science communities due to its lack of quantitative research [6], it has been recognized as basis of further perception research such as behavior, thinking, and pattern recognition. In this paper, we mimic the Gestalt theory framework and model our QA system to think top-down from the question, to mitigate the hallucination issue.

Extending the top-down idea, Schema Theory [5] describes that people build and use schemas as they understand the world. A schema is a pattern or a graph of how concepts of a certain topic relate to each other. It is used to form learnings and reach to solutions when facing a certain problem. Minsky et. al. [20] was the first to bring the Schema concept into computer to develop human-like understanding abilities. Since then many works followed the similar idea and modeled the computational understanding using a graph structure, the most popular of which includes Knowledge Graph [11].

Since the multi-layer perceptron (MLP) [24] was first proposed, there have been several breakthroughs of neural network modeling to increase the understanding capability towards a more generalized AI model solution. Convolutional neural network (CNN) resolves the curse of dimensionality of MLP, and was successfully applied to learn image tasks, a phenomenal example being AlexNet [14]. Generative-Adversarial networks (GAN) [9] first introduced the concept of adversarial training on generative models, which opened up the possibility of generative networks. Most recently, Transformer-based model [26] becomes the most popular choice due to its scalability and flexibility offered by the next-token prediction framework, and is now a foundational building block of most Large Language Models (LLM). However, most of the modern deep models have a bottom-up structure, with no explicit modeling on the global perception. We hypothesize that it is one of the reasons why most LLMs currently suffer from hallucinations, and we introduce the Schema Theory-based framework to mitigate such issue.

2.2 Multimodal Hallucination

Hallucination in multi-modal large language models (MLLMs) refers to the phenomenon where the model generates content that is inconsistent with or unfaithful to the visual inputs [4, 17]. Unlike traditional hallucination issues in language models, hallucination in MLLMs manifests uniquely in the visual-language context, primarily in three aspects: object category errors (identifying non-existent objects), attribute errors (describing incorrect object properties), and relation errors (misrepresenting relationships between objects)[4].

Recent studies have revealed several key factors contributing to hallucination in MLLMs. First, the visual encoder's limitations, such as information loss during encoding and feature bias, can lead to incomplete or inaccurate visual understanding[25]. Second, the strong language priors in large language models may override visual evidence, causing the model to generate responses based on parametric knowledge rather than actual visual content [10]. Third, the cross-modal alignment between visual and textual features remains challenging, where simple connection modules like linear projections may fail to capture complex visual-linguistic relationships[17].

Various approaches have been proposed to mitigate hallucination in MLLMs. Data-centric methods focus on improving training data quality and diversity[16]. Model-centric approaches explore enhanced visual encoders and more sophisticated cross-modal alignment mechanisms[25]. Some researchers have also investigated decoding strategies to maintain visual attention throughout the generation process[12]. However, these methods still struggle with attention misalignment, where models fail to focus on relevant visual regions or miss critical visual details during question answering tasks [10].

Recent benchmarks like POPE[16] and HallusionBench [10] have been developed to systematically evaluate hallucination in MLLMs. These benchmarks reveal that even state-of-the-art models like GPT-4V still face significant challenges in maintaining visual-language consistency, with performance significantly below human level [10]. This gap highlights the need for more effective approaches to address the hallucination problem in MLLMs.

3 Methods

3.1 Schema Theory

Schema theory [5], proposed by psychologist Frederic Bartlett in the 1930s and later expanded by others, explains how people organize knowledge and interpret experiences based on pre-existing mental frameworks or "schemas." Schemas are cognitive structures that help individuals process information efficiently by recognize patterns of behavior, objects, and concepts they have previously encountered. These schemas influence how people perceive, interpret, and remember information. In multimodal perception, schema-based approaches could enable models to understand images and questions more effectively by drawing on contextual or domain-specific knowledge. We use this solution schema as the guidelines.

In summary, as shown in Figure 2, this flow starts by clarifying and fully understanding the problem, ensuring that requirement is clearly identified. Once the problem is comprehensively grasped, the next step is to formulate or recognize the appropriate solution schema to guide the approach. After establishing this framework, the following step is to gather the correct information from reliable

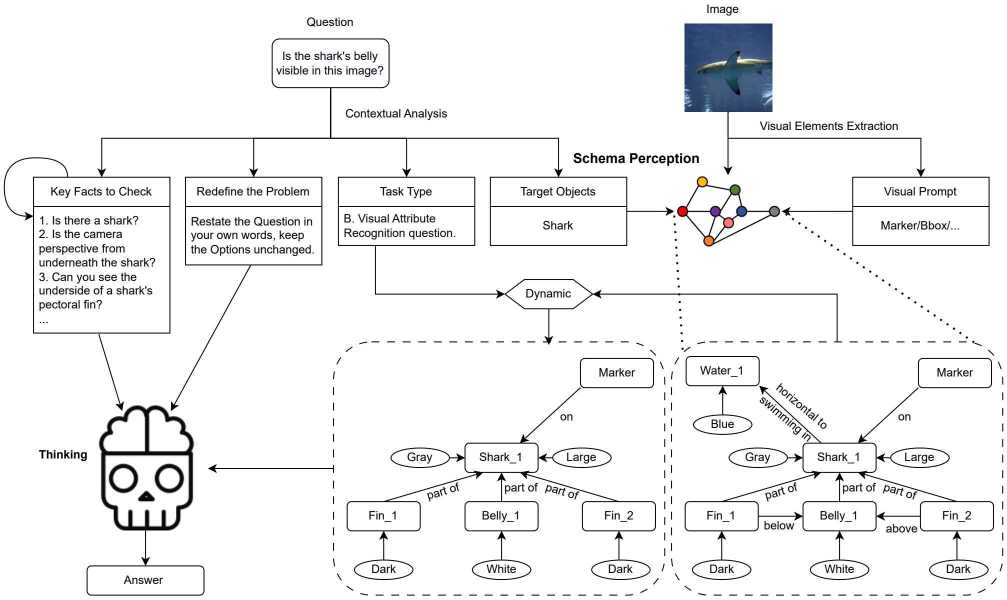

Figure 2: **Work Flow.** This flow starts by clarifying and fully understanding the problem, ensuring that requirement is clearly identified. Once the problem is comprehensively grasped, the next step is to formulate or recognize the appropriate solution schema to guide the approach. After establishing this framework, the following step is to gather the correct information from reliable sources, integrating a strong commitment to factual accuracy. To further reinforce this integrity, fact-checking mechanisms come into play, eliminating any need to fabricate details or rely on incomplete data. By diligently applying these safeguards, the risk of generating hallucination is significantly reduced, leading to a more trustworthy and credible outcome.

sources, integrating a strong commitment to accuracy. To further reinforce this integrity, fact-checking mechanisms come into play, eliminating any need to fabricate details or rely on incomplete data. By diligently applying these safeguards, the risk of generating hallucination is significantly reduced, leading to a more trustworthy and credible outcome.

3.2 Contextual Analysis

For question, we will analyze the question target and the question type. For question target, we extract the key object words we should focus when we analyze the visual content. For question type, based on the question posed about image, the context required to answer them, and the level of reasoning needed. We list 8 VQA categories: Object Recognition question, Visual Attribute Recognition question, Spatial Relationship question, Action Recognition question, Temporal question, Causal question, Predictive question, Comparative question.

3.2.1 Problem digestion

When handling contextual analysis, after identifying the task category, we first ask LLMs to rethink/redefine the question to helps better clarify, contextualize, and refine their understanding of the given question. This step is beneficial for several reasons. First, it can improve Contextual Understanding by helping LLMs anchor their focus to relevant aspects of both the question and the image context. This can reduce ambiguity, especially when a question is vague or could be interpreted in multiple ways. Second, it helps disambiguate complex questions. In some complex questions involving multiple entities or actions, restating clarifies which elements and attributes are being compared, minimizing misinterpretation. Third, it helps align with the task, boosting consistency in responses. When LLMs systematically redefine questions, it helps them maintain consistency across VQA tasks in terms of the similar questions. This makes their answers more reliable and aligned with the question's specific needs. Overall, redefining questions enables LLMs to provide more precise, coherent, and contextually grounded questions, this is very essential for VQA.

3.2.2 Key Facts

Key Facts Check essentially increase the reliability of LVMs by giving an extra layer of validation to the model's interpretation, helping it to be both more precise and reliable in final answer generation. In our method, we have hierarchy design to minimize hallucinations by focusing on grounding responses in observable details, it reduces the likelihood of the model introducing extra, unobserved details into the answer. And this hierarchy design also can offer an early stop mechanism to save cost.

3.3 Visual Analysis

3.3.1 Auto-scaling

For general usage, a moderate resolution that retains essential details is often ideal. If the question requires complex or small details, a higher resolution can improve accuracy, but if the question asks simple or has large and easily recognizable elements in a simple image, a lower resolution may suffice. Furthermore, larger images take more time and cost to process, so providing an excessively high-resolution image could slow down the task without significantly improving accuracy.

Although many models include preprocessing steps that scales high resolution image inputs to not exceed a specific range. If your input image is significantly higher than this specific resolution, the preprocessing will scale the image to the maximum size, leading to potential cost waste. If your input image is excessively low-resolution image, the performance may be greatly reduced.

We conduct auto-scaling on images before using them for final tasks. We input image and question to LLM-autoscaler (adapt to the training sizes while keep the resolution ratio) to decide the input resolution. This process will maintain consistency with the model's expected input dimensions, optimizing computational efficiency, and ensuring accurate feature extraction.

3.3.2 Visual Anchor

Vision encoders in the CLIP family, including BLIP series, are primarily designed for semantic visual representations but are challenged by ambiguous encoding. This further leading to the lack of spatial information for MLLMs to localize the target objects, which weakens the reasoning ability of language model part and further increase the hallucination. However, we value the semantic representations, but how can we increase the grounding ability but not change the visual encoder dramatically.

We use a simple marker as visual anchor to simplify the language model part's process of identifying relevant target within the visual content. By associating specific parts of the image with discrete markers, the model can better match text-based prompts to visual regions, improving both accuracy and response specificity. This anchor is optional in our framework.

3.4 SoT Perception

Below is a step-by-step description of how the Schema of Thought (SoT) approach can be applied to reduce hallucination in Large Multimodal Language Models (MLLMs) by leveraging Schema Theory. The approach focuses on leveraging the above extracted information, aligning them with the task schema, and ultimately ensuring the final answer is grounded in verified, factual information.

Identify the Schema: The first step is to clearly define the question. We ask LLMs to classify the question into our predefined 9 schemas (8 VQA categories + 1 “Not belong to any of them" category). Then beyond this background, we ask the model to identify the core intent of the user's query (the “question target”) and redefine the problem in this schema. This processing narrows down what type of information or reasoning is necessary. And this help interpret and reshape the query information based on the cognitive structures, or schemas.

Generate a SoT Graph: Base on the question target, input visual information, the visual anchor (if we have) to generate a scene graph. If available, visual anchors (e.g., marker or bounding boxes) provide an initial spatial representation in the visual information. These anchors help the model locate specific objects or regions that are relevant to the question. For the graph, we will prune this graph to be a task-schema—essential structure that specifies which elements of the scene are most relevant. This helps eliminate extraneous details from the scene graph that do not directly pertain to the question. By focusing on the refined question, we remove nodes and edges in the scene graph that don't serve the primary query. This step mirrors the cognitive process of filtering out irrelevant stimuli and focusing on the task schema.

Final Answer: In this final step, the model integrates the redefined question (focused on a clear goal), the pruned graph (with only relevant objects, attributes, and relationships), the verified key facts. This integrated approach encourages the model to produce an answer that is (a) aligned with the user's query, (b) grounded in the visible evidence from the scene, and (c) validated through fact-checking.

By applying the Schema of Thought approach in MLLMs, we leverage a cognitive-psychology-inspired structure to mitigate hallucinations. we refine the user task in accordance with human cognitive schemas and validate crucial information through Key Facts Check. This process, grounded in structured reasoning and verification, ultimately reduces the model's tendency to fabricate details and ensures that the final answer is credible, relevant, and factually aligned.

4 Experiments

Method	Prompt	Evaluation	Performance Acc(%)
Human	/	Human Check Description	95.7
ChatGPT-4o	/	Human Check Description	68.0
ChatGPT-4o (Ours)	SoT Perception	Human Check Description	98.0
Llama3.2-11B	/	Automatic	37.3
Llama3.2-11B (Ours)	SoT Perception	Automatic	42.0
ChatGPT-4o	/	Automatic	64.7
ChatGPT-4o (Ours)	SoT Perception	Automatic	78.0
Claude3.5	/	Automatic	31.3
Claude3.5 (Ours)	SoT Perception	Automatic	59.3

Table 1: Comparison on MMVP (Eyes Wide Shut). For human check, we check the SoT graph and the output description manually. From Our experiments, ChatGPT-4o with SoT achieved 98% performance with 30% improvement over the original ChatGPT-4o (68%). And it is the first time MLLM beats human (95.7%) on visual question task. This demonstrated that SoT has a strong theoretical upper bound. For automatic manner, we directly use SoT generated by Claude 3.5 Sonnet. We tested both open-source and commercial VLMs. For open-source Llama 3.2, SoT achieved 4.7% improvement. For commercial VLMs, ChatGPT-4o with SoT achieved 13.3% improvement, Claude 3.5 Sonnet with SoT achieved 28% improvement. This shows that our SoT can benefit both open-source and commercial models.

We perform experiments on three datasets including MMVP, HallusionBench, and AutoHallusion with both open-source Large Vision-Language Models (Llama 3.2 Vision) and commercial Large Vision-Language Models (Claude 3.5, ChatGPT-4o). We assess performance by calculating the accuracy of correct answers in a manner of visual question answering.

4.1 Benchmarks

MMVP [25] proposed CLIP-blind pairs, images that CLIP perceives as similar despite their clear visual differences, demonstrating CLIP-based vision encoder struggles to encode "properly". The MMVP benchmark is designed to systematically evaluate the performance of recent CLIP-based models in understanding and processing visual patterns. It distills a subset of questions from the original MMVP benchmark into simpler language descriptions, categorizing them into distinct visual patterns. Each visual pattern is represented by 15 text-image pairs. The benchmark assesses whether CLIP models can accurately match these image-text combinations, providing insights into the capabilities and limitations of these models.

HallusionBench [10] proposed that the strong language prior in most of the existing SOTA LVLMs can be a double-edged sword: they may ignore the image context and solely rely on the (even contradictory) language prior for reasoning. In contrast, the vision modules in VLMs are weaker than

Method	Evaluation	Question Pair	Figure	Easy	Hard	All
		(*qAcc*) ↑	(*fAcc*) ↑	(Easy *aAcc*) ↑	(Hard *aAcc*) ↑	(*aAcc*) ↑
GPT-4 (CVPR2024)	GPT4-Assisted	28.79	39.88	75.6	37.7	65.3
Claude3 (CVPR2024)	GPT4-Assisted	21.76	28.61	55.16	41.4	56.86
ChatGPT-4o	GPT4-Assisted	37.80	47.40	78.46	46.28	68.64
ChatGPT-4o-SoT (Ours)	GPT4-Assisted	46.15	52.02	79.34	54.42	71.83
Claude3.5	GPT4-Assisted	38.46	47.4	69.89	54.18	68.82
Claude3.5-SoT (Ours)	GPT4-Assisted	42.20	46.25	73.19	57.44	71.66
Lamma3.2-11B	GPT4-Assisted	10.99	15.90	27.91	27.91	36.67
Lamma3.2-11B-SoT (Ours)	GPT4-Assisted	15.60	17.92	36.04	36.05	44.82

Table 2: **Correctness Leaderboard on HallusionBench with various MLLMs:** For commercial model, Claude 3.5 Sonnet with SoT achieved 2.84% - 3.74% improvement. ChatGPT-4o with SoT achieved 3.19% - 8.35% improvement. For open-source model, Llama 3.2 with SoT achieved 2.02% - 8.59% improvement. And our SoT significantly improves accuracy on hard problems.

Method	Quality	Evaluation	Synthetic	Synthetic SR	Real-World	Real-World SR	All
			(*sAcc*) ↑	(*ssrAcc*) ↑	(*rAcc*) ↑	(*rsrAcc*) ↑	(*aAcc*) ↑
ChatGPT-4o	1,2,3	GPT4	75.42	68.44	69.86	58.83	72.93
ChatGPT-4o-SoT (Ours)	1,2,3	GPT4	77.68	70.94	72.64	62.90	75.42
ChatGPT-4o	2,3	GPT4	78.87	71.43	74.00	64.60	76.84
ChatGPT-4o-SoT (Ours)	2,3	GPT4	80.65	76.27	74.65	66.80	78.15
ChatGPT-4o	3	GPT4	76.96	76.67	76.34	71.54	76.60
ChatGPT-4o-SoT (Ours)	3	GPT4	79.58	82.22	78.24	73.08	78.81
Claude3.5	1,2,3	GPT4	59.63	39.22	49.21	27.19	54.97
Claude3.5-SoT (Ours)	1,2,3	GPT4	63.16	50.31	54.29	37.61	59.19
Claude3.5	2,3	GPT4	66.15	47.47	54.15	31.40	61.15
Claude3.5-SoT (Ours)	2,3	GPT4	68.47	54.84	58.14	41.60	64.16
Claude3.5	3	GPT4	73.30	72.22	65.27	52.31	68.65
Claude3.5-SoT (Ours)	3	GPT4	76.96	78.89	65.27	64.62	70.20
Lamma3.2-11B	1,2,3	GPT4	49.22	35.16	42.50	28.08	46.21
Lamma3.2-11B-SoT (Ours)	1,2,3	GPT4	60.21	43.75	59.71	46.89	59.99
Lamma3.2-11B	2,3	GPT4	52.12	32.26	44.88	27.4	49.10
Lamma3.2-11B-SoT (Ours)	2,3	GPT4	64.66	44.70	62.57	48.6	63.80
Lamma3.2-11B	3	GPT4	43.98	15.56	44.66	26.15	44.37
Lamma3.2-11B-SoT (Ours)	3	GPT4	60.21	46.67	63.74	46.92	62.25

Table 3: **Correctness Leaderboard on AutoHallusion with various MLLMs:** For different quality settings, different scene settings, different models, SoT can consistently improve the performance. Especially, our SoT significantly improves the Synthetic Spatial Relation performance, the highest improvement with our SoT can reach 31.11%, demonstrating the promising of our proposed SoT. SR: Spatial Relation

LLMs and may result in misleading visual representations, which are then translated to confident mistakes by LLMs.

AutoHallusion [30] found that certain context cues in an image may trigger the language module's overconfident and incorrect reasoning on abnormal or hypothetical objects. Though a few benchmarks have been developed to investigate LVLM hallucinations, they mainly rely on hand-crafted corner cases whose fail patterns may hardly generalize, and finetuning on them could undermine their validity. AutoHallusion probes the language modules in LVLMs for context cues and uses them to synthesize images by: (1) adding objects abnormal to the context cues; (2) for two co-occurring objects, keeping one and excluding the other; or (3) removing objects closely tied to the context cues. It then generates image-based questions whose ground-truth answers contradict the language module's prior. A model has to overcome contextual biases and distractions to reach correct answers, while incorrect or inconsistent answers indicate hallucinations.

4.2 Multimodal Large Language Models

To evaluate our Gestalt Perception, we conduct experiments on both open-source SOTA Multimodal Large Language Models (Llama 3.2 Vision [1]) and SOTA commercial Multimodal Large Language Models (Claude 3.5 [2], ChatGPT-4o [23]).

Llama 3.2 Vision [1] supports image reasoning use cases, such as document-level understanding including charts and graphs, captioning of images, and visual grounding tasks such as directionally pinpointing objects in images based on natural language descriptions. Llama 3.2 can also bridge the gap between vision and language by extracting details from an image, understanding the scene, and then crafting a sentence or two that could be used as an image caption to help tell the story. We use 11B version for all experiments.

Claude 3.5 Sonnet [2] reaches graduate-level reasoning (GPQA), undergraduate-level knowledge (MMLU), and coding proficiency (HumanEval). It shows marked improvement in grasping nuance, humor, and complex instructions, and is exceptional at writing high-quality content with a natural, relatable tone.

GPT-4o [23] aims toward much more reliable for various tasks, it accepts as input any combination of text, audio, image, and video and generates any combination of text, audio, and image outputs. It can respond to audio inputs in as little as 232 milliseconds, with an average of 320 milliseconds, which is similar to human response time in a conversation. It matches GPT-4 Turbo performance on text in English and code, with significant improvement on text in non-English languages, while also being much faster. GPT-4o is especially better at vision and audio understanding compared to existing models.

4.3 Results

On MMVP, as show in Table 1, we conduct different evaluations including human check and automatic manner. For human check, we check the SoT graph and the output description manually. From Our experiments, ChatGPT-4o with SoT achieved 98% performance with 30% improvement over the original ChatGPT-4o (68%). And it is the first time MLLM beats human (95.7%) on visual question task. This demonstrated that SoT has a strong theoretical upper bound. For automatic manner, we directly use SoT generated by Claude 3.5 Sonnet. We tested both open-source and commercial VLMs. For open-source Llama 3.2, SoT achieved 4.7% improvement. For commercial VLMs, ChatGPT-4o with SoT achieved 13.3% improvement, Claude 3.5 Sonnet with SoT achieved 28% improvement. This shows that our SoT can benefit both open-source and commercial models.

On Hallusionbench, as show in Table 2, for commercial model, Claude 3.5 Sonnet with SoT achieved 3.74% improvement on Question Pair Accuracy, 3.3% improvement on Easy Case Accuracy, 3.26% on Hard Case Accuracy, 2.84% improvement on All Accuracy. ChatGPT-4o with SoT achieved 8.35% improvement on Question Pair Accuracy, 4.62% improvement on Figure Accuracy, 8.14% on Hard Case Accuracy, 3.19% improvement on All Accuracy. For open-source model, Llama 3.2 with SoT achieved 4.31% improvement on Question Pair Accuracy, 2.02% improvement on Figure Accuracy, 8.31% improvement on Easy Case Accuracy, 8.59% on Hard Case Accuracy, 8.51% improvement on All Accuracy. From our experiments, our SoT significantly improves accuracy on hard problems.

On AutoHallusion, as show in Table 3, for different quality settings (1 means low quality, 2 means medium quality, 3 means high quality), different scene settings, different models, SoT can consistently improve the performance. Especially, our SoT significantly improves the Synthetic Spatial Relation performance, the highest improvement with our SoT can reach 31.11%, demonstrating the promising of our method.

5 Conclusion

In conclusion, our Schema of Thought (SoT) approach leverages human cognitive psychology principles to reduce hallucination in multimodal large language models and improve performance on visually challenging question-answering tasks. By employing Schema Theory, SoT enables more accurate processing of multimodal input, as it taps into systematical framework to integrate and interpret text/visual cues with significantly reduced cognitive load. The experimental results on multiple benchmark datasets—including Eyes Wide Shut (MMVP), HallusionBench, and Auto-Hallusion—demonstrate that SoT not only mitigates hallucination but also streamlines information organization for enhanced accuracy. Our findings highlight the promise of SoT as a robust and grounded prompting technique, paving the way for more reliable visual question-answering models.

References

[1] Meta AI. Llama 3.2, 2024. Large language model.

[2] Anthropic. Claude 3.5, 2024. Large language model.

[3] Michael A Arbib. Schema theory. *The encyclopedia of artificial intelligence*, 2:1427–1443, 1992.

[4] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[5] Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1995.

[6] Vicki Bruce, Mark A Georgeson, and Patrick R Green. *Visual perception: Physiology, psychology and ecology*. Psychology Press, 2014.

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[8] Michael W Eysenck and Marc Brysbaert. *Fundamentals of cognition*. Routledge, 2018.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[11] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

[12] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[13] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

[16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[17] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[18] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*, 2021.

[19] Mary B McVee, Kailonnie Dunsmore, and James R Gavelek. Schema theory revisited. *Review of educational research*, 75(4):531–566, 2005.

[20] Marvin Minsky et al. A framework for representing knowledge, 1974.

[21] Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. How additional knowledge can improve natural language commonsense question answering? *arXiv preprint arXiv:1909.08855*, 2019.

[22] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024.

[23] OpenAI. Chatgpt-4o, 2024. Large language model.

[24] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[25] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[26] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[27] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[28] Max Wertheimer. Experimentelle studien uber das sehen von bewegung. *Zeitschrift fur psychologie*, 61:161–165, 1912.

[29] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[30] Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *arXiv preprint arXiv:2406.10900*, 2024.

[31] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.

[33] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *arXiv preprint arXiv:2406.09136*, 2024.