Reducing Hallucinations of Large Vision Language Models

Zhanyang Gong andygong520@g.ucla.edu
Agustin Costa agustincosta@g.ucla.edu
Varun Agarwal varunagarwal1@g.ucla.edu

Abstract

Large Vision-Language Models (LVLMs) are capable of understanding both images and text, enabling impressive performance on multimodal tasks. However, these models remain susceptible to produce hallucinations that can undermine their reliability, particularly in high-stakes scenarios. This project systematically investigates and compares two complementary strategies for hallucination reduction: (1) visual grounding by integrating state-of-the-art object detectors (YOLOv8 and DINO-X) to supply explicit image-derived context, and (2) reasoning-enhanced language prompting by applying structured prompting techniques such as Chain of Thought and Chain of Verification to improve reasoning and self-correction. Our experiments leverage two benchmarks—the POPE adversarial object detection dataset and the Visual Puzzles reasoning dataset—to evaluate the impact of each approach on accuracy, hallucination rate, and related metrics across multiple open-source models. Results show that vision detection helps to reduce object-related hallucinations and improves factual alignment, while reasoning-based prompting offers notable gains on complex, logic-driven tasks. Importantly, we find that the efficacy of each strategy is highly dependent on both the underlying model and the task type, suggesting that a hybrid approach may provide the most robust solution.

1 Introduction to the Problem

Large Vision-Language Models (LVLMs), which are capable of understanding both visual and textual data, have demonstrated impressive capabilities on multi-modal tasks. However, they often suffer from hallucination-the generation of confident yet factually incorrect outputs, which is a phenomenon that our project will aim to systematically analyze and reduce.

The current LVLMs shortcomings appear to have a wide variety of potential solutions, allowing for a wide range of learning opportunities. We chose this topic to explore and learn about, starting from data selection to the method of hallucination reduction, and finally the evaluation methods. As a group, we have a common interest in state-of-the-art ML models and want to work with such models.

2 Literature Review and Baseline Model Selection

Bai et al. (2024) reviewed the most recent improvement of models that reduces hallucinations, and found that current hallucination reduction techniques have seen notable progress that has translated into stronger models. Techniques such as instruction tuning (Liu et al., 2023; Krishna et al., 2017), reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Yu et al., 2024), and contrastive learning (Jiang et al., 2024; Sarkar et al., 2024) have been effective in grounding model outputs more firmly in visual input. However, challenges remain, particularly in complex visual contexts or abstract reasoning, where hallucinations persist, highlighting the need for more robust evaluation benchmarks and model interpretability tools(Bai et al., 2024).

The model selection process started with a review of existing Large Vision-Language Models (LVLMs). A common evaluation dataset was chosen to allow for a fair comparison across all selected models. The selection was limited to open-source models with fewer than 10 billion parameters. This constraint was set to ensure the models could run on standard GPU instances with 16GB of memory, such as the T4 available on Google Colab. While larger and more advanced models will be included later in the project as benchmarks, we initially focused on smaller models. This is because high-performing models may already exhibit strong capabilities, leaving less room for evaluating the impact of hallucination-reduction techniques.

Model	Accuracy
LLaVA 1.5 7B	26%
Gemma 3 4B	28%
SmolVLM 2B	23%
Mini InternVL 1.5 2B	39%
InternVL 2.5 4B	32%
Qwen VL 7B (8-bit)	33%

Table 1: summarization of the results of running a preliminary evaluation using 100 samples from the Visual Puzzles dataset

To identify suitable models, resources such as OpenGVLab's MultiModality Arena and Hugging Face's model repository (filtered by Visual Question Answering tasks) were used. These platforms provided valuable insights into models that met our criteria and allowed for a comparison of their performance and popularity.

The Visual Puzzles dataset(Song et al., 2025), developed by researchers at Carnegie Mellon University, was chosen for evaluation. This dataset presents complex visual tasks that go beyond simple object detection and require reasoning based on image content. Each puzzle includes an image, a prompt, and four possible answer choices. Notably, even top-performing proprietary models, such as OpenAI's o1, have achieved no more than 51% accuracy on this dataset. This highlights the challenge it presents and its potential for revealing hallucination-related weaknesses in LVLMs. Additionally, we believe that the hallucination mitigation methods developed in this study may be particularly effective on this dataset.

Table 1 summarizes the results of running a preliminary evaluation using 100 samples from the Visual Puzzles dataset.

While the performance across models is relatively close, we have chosen LLaVA 1.5 7B (Liu et al., 2024) as the primary model for this study. This decision is based on its widespread use in the LVLM research community and its presence in most major benchmarks. LLaVA 1.5 is built on two well-established architectures, Vicuna (Zheng et al., 2023) and CLIP (Radford et al., 2021), which adds to its credibility and suitability for in-depth analysis. Although its initial accuracy is not the highest, we believe it offers a strong foundation for testing and improving hallucination-reduction techniques.

3 Approaches

Two enhancement strategies are applied: (1) Vision-based methods, and (2) Language-based methods.

The outputs of the enhanced models are compared to the original base model through accuracy and hallucination event analysis. The evaluation results are then presented through a Gradio (Abid et al., 2019) web app, and the improved models are saved for future use or deployment.

3.1 Vision Based Methods

The first methodology involves integrating an additional vision backbone model to enhance LVLMs with specific capabilities through object detection models. Traditional deep object detectors often fail to effectively capture contextual information, including visual scenes and human language (Zang et al., 2025). To address these limitations, several recent studies have proposed two-step approaches such as: (1) the generate-then-detect framework (Zang et al., 2025); (2) generating natural language responses intertwined with segmentation masks; (3) decomposing regions of interest to generate corresponding responses (Ma et al., 2024); and (4) employing a mask decoder for language generation and reasoning segmentation (Lai et al., 2024).

Despite these advancements, few studies have specifically focused on reducing hallucinations in LVLMs and systematically comparing their performance. Building on insights from existing work, this part of the project will integrate a segmentation-based vision model backbone with LVLM base models to investigate the impact of this enhancement.

We implemented an object-aware prompting strategy using two object detection models—YOLOv8 (Reis et al., 2024) and DINO-X (Ren et al., 2025)—as visual preprocessors. The goal of this approach is to extract explicit object-level information from each image and inject it into the LVLM's prompt to anchor its output to finetuned visual facts.

YOLOv8x was selected for its speed and efficiency in detecting common object categories. It is a convolutional-based model meant for real-time inference and is capable of producing accurate bounding box outputs across a broad range of predefined classes. The v8x model is the largest in the v8 series. In contrast, DINO-X is a transformer-based detector trained on large-scale grounding datasets with open-vocabulary coverage. Its key strength

lies in prompt-free open-world detection, allowing it to identify rare, ambiguous, or relational objects without needing predefined class labels. We expect the YOLOv8 model to perform better on images with clearly defined, known objects while DINO-X should outperform on abstract and more difficult images.

In our workflow, each input image was first processed by one of the object detection models to extract a list of detected objects. These detections were converted into structured textual descriptions (e.g., "Objects detected in the image: dog, skateboard, traffic cone") and prepended to the original user prompt. The combined prompt was then passed to the base LVLM (LLaVA), conditioning its response on verified visual content. This augmentation helps mitigate hallucination by grounding the model's reasoning in elements that are actually present in the image.

3.2 Language Based Methods

The second methodology for mitigating hallucinations focuses on the application of prompting and language-based strategies directly to LVLMs. Proposed approaches include Chain of Thought (CoT) prompting, requerying, self-verification, and combinations of these techniques. Prior studies, such as Woodpecker (Yin et al., 2024), introduced self-verification mechanisms to assess the factual consistency of generated outputs, while others, including reasoning frameworks (Wu et al., 2025), were specifically developed to address hallucination phenomena. Drawing on these advances, we systematically applied individual techniques and their combinations to one or more base models to assess and compare their effectiveness.

3.2.1 Chain of Thought (CoT) Prompting

Chain of Thought prompting has emerged as an effective technique for enhancing reasoning in large language models (Wei et al., 2023). For our implementation, we structure reasoning prompts to the base model to generate specific types of reasoning according to the dataset being evaluated.

Each strategy incorporates a standardized answer format instruction that clearly separates the reasoning process from the final answer, making evaluation more consistent and reliable, as suggested by Kojima et al. (2023).

3.2.2 Chain of Verification (CoVe) Prompting

Chain of Verification implements a two-stage reasoning process where the model first generates an initial answer with reasoning, then critically examines its own response. This approach draws from Woodpecker framework and extends it with visual reasoning capabilities. This approach allows the model to catch and correct its own mistakes, particularly in cases where the initial reasoning contained logical errors or missed critical visual elements, complementing the approach of Wu et al. (2025) for reducing hallucinations through iterative reasoning.

Our implementations of both techniques are evaluated against direct responses (without enhanced reasoning prompts) to measure their effectiveness in reducing hallucinations and improving accuracy on visual reasoning tasks.

3.3 Dataset and Evaluation Framework

The hallucination reduction techniques implemented in this project target different types of weaknesses in LVLMs, each addressing specific hallucination sources through complementary approaches.

Object Detection Augmentation vs. Reasoning Enhancement

Using a separate vision model as a preprocessing step primarily benefits datasets focused on object detection evaluation, such as the Pooling-based Object Probing Evaluation (POPE) (Li et al., 2023). This technique enriches the LVLM's understanding by (1) detecting objects present in the image with high precision, (2) providing richer and more complete context to the LVLM, and (3) conditioning the model to generate more accurate outputs by grounding its responses in detected visual entities

For the specific case of POPE, which tests a model's tendency to hallucinate non-existent objects by asking yes or no questions about the presence of objects in the image, the vision model preprocessing makes the LVLM's task substantially more straightforward by directly supplying object presence information.

In contrast, the VisualPuzzles dataset (Song et al., 2025) evaluates models on tasks requiring complex visual reasoning and problem-solving skills that go beyond simple object detection. These tasks benefit from language-based techniques that enhance the model's reasoning process rather than its perceptual abilities. The dataset contains a balanced combination of inductive, deductive, spatial,

analogical and algorithmic puzzles with 4 possible options to choose from.

Single-pass vs. Multi-step Reasoning

With a direct (single-pass) approach, the model generates a response in one forward pass, typically outputting just the option letter corresponding to the correct answer. This approach has significant limitations for complex reasoning tasks: the model must perform the entire reasoning chain internally without explicitly articulating intermediate steps, all deductions must occur simultaneously within a single forward pass, and the model cannot refine or correct its thinking process after initial consideration

Our language-based approaches address these limitations through structured multi-step reasoning to break down complex visual tasks into manageable steps, make explicit intermediate deductions, build toward a final conclusion through progressive reasoning and identify and correct potential errors or oversights.

Our evaluation approach was designed to provide a rigorous assessment of the various hallucination reduction methods across two distinct task types. For each method, a fixed number of randomized samples were chosen from the respective datasets, with controlled seed settings to ensure reproducibility and fair comparison across different approaches.

Experimental Setup

All evaluations were conducted on a Google Cloud Platform (GCP) virtual machine equipped with an NVIDIA Tesla T4 GPU with 16GB VRAM. Models were run with FP16 precision when possible to optimize memory usage while maintaining inference quality. For both the POPE and Visual Puzzles datasets, we maintained consistency in testing conditions to ensure valid comparisons. Dataset samples were randomly selected but with fixed seeds to ensure reproducibility while covering a diverse range of test cases.

Each model was configured with consistent system prompts tailored to the specific task domain (object detection for POPE, visual reasoning for Visual Puzzles).

Evaluation Metrics

We tracked several key metrics to comprehensively assess model performance:

- Accuracy: The proportion of correct answers among all evaluated samples.
- Hallucination Rate: The rate of falsely identifying non-existent objects.
- Miss Rate: The rate of failing to identify present objects in POPE (false negatives).
- Precision and Recall: Particularly relevant for POPE, these metrics provided deeper insights into the nature of errors.

The results of these evaluations were captured in comprehensive CSV files recording the full model responses, extracted answers, and correctness assessments for subsequent analysis. These files enabled detailed post-hoc analysis and visualization of performance patterns across models and reasoning strategies.

This standardized evaluation framework allowed us to systematically compare the effectiveness of different hallucination reduction techniques across diverse visual reasoning tasks, revealing important insights about their task-specific efficacy and limitations.

4 Experimental Results

4.1 Vision Object Detection

In this part of the research, we experimented with both YOLOv8 and DINO-X as object detection backbones to enhance grounding in LVLMs. Both models aim to reduce hallucination by supplying explicit visual context, but they differ in architecture and detection philosophy. YOLOv8 is a fast, cheap, high-accuracy, convolutional-based detector optimized for real-time object detection. It is widely used for its efficiency and precision in detecting common object categories. In contrast, DINO-X is a transformer-based encoder-decoder model trained on large-scale grounding datasets with open-vocabulary coverage, enabling zero-shot detection of diverse and rare object types in images and videos. A key feature of DINO-X is its unique capability for prompt-free "anything" detection; it can identify ambiguous, relational, or uncommon objects without needing explicit prompt guidance.

In our proposed workflow, the input image is first analyzed by either YOLOv8 or DINO-X to extract a list of detected objects. These object labels are then transformed into a structured text format and combined with the original user prompt before being fed into the base LVLM (e.g., LLaVA). This augmentation allows the LVLM to ground its reasoning in visual facts. While DINO-X also outputs

Table 2: Performance comparison of LLaVA, DINO-X-enhanced LLaVA, and YOLO-enhanced LLaVA on object grounding tasks. Metrics include: TP (True Positives), FP (False Positives), TN (True Negatives), FN (False Negatives), Acc. (Accuracy), Pre. (Precision), Rec. (Recall), and F1 (F1 Score).

Metric	LlaVA	DINO-X	YOLO
TP	539	554	1282
FP	208	211	128
TN	392	389	1373
FN	61	46	216
Acc.	77.6%	78.6%	88.5%
Pre.	72.2%	72.4%	90.9%
Rec.	89.8%	92.3%	85.6%
F1	80.0%	81.2%	88.2%
Total Images	200	200	500

grounded visual overlays, these were used only for qualitative analysis and not included as model inputs.

We compared both object detectors on the POPE dataset to assess their respective impact on hallucination rates, analyzing which method more effectively anchored the LVLM's output to the actual image content.

Both qualitative and quantitative assessments were conducted to find the effectiveness of using object detection in LVLM prompts. The DINO-X evaluation was limited to approximately 200 images from the COCO dataset due to rate limits and cost. For each image, a set of positive and adversarial questions were paired, resulting in a total of 1,200 evaluation prompts. Object detection outputs from DINO-X were formatted into text and appended to each prompt before being passed into LLaVA. Separately, YOLOv8 was used to generate object tags for a larger subset of 500 COCO images, following the same augmentation and evaluation procedure. The outputs from both enhanced pipelines—DINO-X + LLaVA and YOLOv8 + LLaVA—were compared against the baseline performance of vanilla LLaVA (no object grounding). The results are summarized in the Table 2.

The enhanced prompts consistently outperformed the baseline. DINO-X + LLaVA showed gains across all metrics, particularly in recall (+2.5%), indicating better identification of relevant image-grounded content. YOLOv8 + LLaVA, despite being a more lightweight approach, yielded significant improvements in accuracy, precision, and F1—suggesting that even coarse object grounding with a fast detector like YOLOv8 can substantially improve factual alignment and reduce hallucinations. Both approaches helped reduce references to non-existent objects. However, DINO-X appeared more capable of identifying abstract or rare object types (e.g., "person holding a reflective surface"), while YOLOv8 contributed stronger results on common object detection due to its real-time optimization and broader image coverage.

In the qualitative analysis, a Gradio App was developed to compare the outputs of the vanilla LLaVA model and the LLaVA model enhanced with DINO-X. It was observed that integrating the object detector helped correct ambiguous object references, allowing the LVLM to generate more accurate responses. For instance, there is an example (1) where the prompt asked whether skis were present in the image. The vanilla LLaVA incorrectly responded "yes" because it detected a skier. However, the ground truth was that the person was riding a snowboard, meaning no skis were actually present. With DINO-X, the model correctly recognized the object as a snowboard, reducing hallucination and aligning better with the visual evidence. Instead of relying solely on high-level context or co-occurrence patterns (e.g., associating a skier with skis), the enhanced model can disambiguate visually similar or contextually misleading elements, leading to more faithful and accurate image-text alignment.

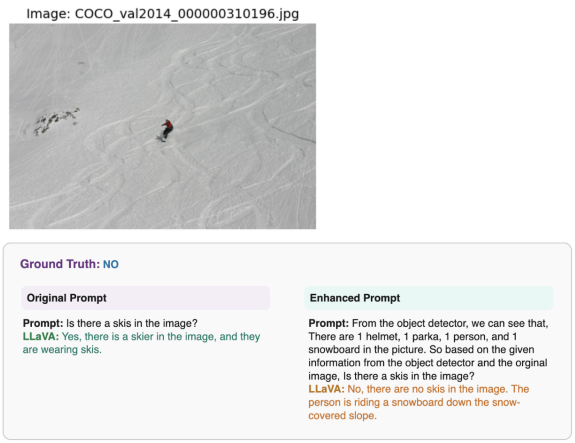

Figure 1: An example of DINO-X LLaVA model that revealed capability of identify ambiguous objects and correct the original model output by using enhanced object detection prompt.

4.2 Language-Based Methods: Chain of Thought and Chain of Verification

Our experiments tested our base model (LlaVA-1.5-7B) and two additional models: Gemma-3-4B-it (Google) and GPT-4.1-mini (OpenAI), using different reasoning strategies: direct answering, Chain of Thought (CoT), and Chain of Verification (CoVe). For CoVe in particular, we track both the response from the initial reasoning and after the verification.

4.2.1 Object Detection Results (POPE Dataset)

The POPE dataset evaluation focused on testing models' ability to correctly identify objects present in images while avoiding hallucination of non-existent objects.

Table 3: Detailed evaluation of LLaVA-1.5-7B, Gemma-3-4B-IT, and GPT-4.1-Mini on the POPE dataset, including direct and Chain of Thought accuracy, hallucination, and classification metrics.

Metric	LLaVA	Gemma	GPT-4.1
Acc. (Direct)	0.835	0.825	0.560
Acc. (CoT)	0.760	0.825	0.825
Hal. Rate	0.035	0.055	0.020
Miss Rate	0.205	0.120	0.155
Precision	0.965	0.945	0.980
Recall	0.795	0.880	0.845
F1 Score	0.872	0.911	0.908

Table 4: Step-wise performance of LLaVA-1.5-7B, Gemma-3-4B-IT, and GPT-4.1-Mini on the POPE dataset, including initial reasoning, verification accuracy, and classification metrics.

Metric	LLaVA	Gemma	GPT-4.1
Acc. (Direct)	0.835	0.825	0.560
Acc. (Reas.)	0.820	0.835	0.820
Acc. (Verf.)	0.690	0.835	0.780
Hal. Rate	0.010	0.045	0.020
Miss Rate	0.115	0.105	0.100
Precision	0.990	0.955	0.980
Recall	0.885	0.895	0.900
F1 Score	0.935	0.924	0.938

The POPE dataset results reveal model-specific responses to reasoning enhancement strategies. Gemma-3-4B-it demonstrated remarkable consistency (0.825-0.835 accuracy) across all approaches, suggesting inherent robustness in object recognition tasks. GPT-4.1-mini showed the most dramatic improvement with structured reasoning, increasing from a relatively weak direct performance (0.560) to competitive accuracy with Chain of Thought (0.825) - a substantial 26.5% improvement. Contrary to expectations, LLaVA-1.5-7B performed best with direct answering (0.835), with Chain of Verification actually degrading its performance to 0.690. This pattern suggests that while reasoning enhancements significantly benefit some models (particularly GPT-4.1-mini), they may interfere with others' already-optimized object recognition capabilities. Notably, all models maintained high precision (>0.945) across strategies, indicating strong resistance to hallucinating non-existent objects regardless of the reasoning approach employed.

Self-verification showed a significant 22% improvement for GPT-4.1-mini compared to direct answering. Chain of Thought provided a 26.5% boost for GPT-4.1-mini. Interestingly, LLaVA experienced a performance decrease with both strategies.

4.2.2 Visual Reasoning Results (VisualPuzzles Dataset)

The Visual Puzzles dataset presented a more challenging test of models' visual reasoning capabilities, requiring complex pattern recognition rather than simple object detection.

Table 5: Comparison of three vision-language models—LLaVA-1.5-7B, Gemma-3-4B-IT, and GPT-4.1-Mini—on visual reasoning tasks on Visual Puzzles dataset using Chain of Thought

Metric	LLaVA	Gemma	GPT-4.1
Direct Acc.	0.160	0.255	0.305
Accuracy	0.210	0.235	0.320
Hal. Rate	0.790	0.765	0.680

Table 6: Comparison of three vision-language models—LLaVA-1.5-7B, Gemma-3-4B-IT, and GPT-4.1-Mini—on visual reasoning tasks on Visual Puzzles dataset without Chain of Verification

Metric	LLaVA	Gemma	GPT-4.1
Direct Acc.	0.160	0.255	0.305
Accuracy	0.165	0.200	0.075
Hal. Rate	0.835	0.800	0.925

The Visual Puzzles dataset results reveal surprising limitations of reasoning enhancement strategies on complex visual tasks. GPT-4.1-mini achieved the best baseline performance (0.305 direct accu-

racy), with Chain of Thought providing a modest improvement to 0.320. However, self-verification dramatically degraded its performance to just 0.075 - a 75% reduction. This counterintuitive finding suggests that complex verification processes can interfere with the model's initial visual reasoning capabilities, potentially introducing overthinking or confidence reduction. LLaVA-1.5-7B showed a moderate benefit from Chain of Thought (0.160 to 0.210), while Gemma-3-4B-it saw minimal improvements. In particular, all models struggled with this challenging dataset, showing high hallucination rates (0.680) regardless of strategy. These results indicate that while structured reasoning can help models with simple object recognition tasks, it may be insufficient or even detrimental for complex visual puzzles requiring sophisticated pattern recognition and abstract reasoning.

5 Conclusions and Future Works

In this project, we explored two complementary strategies to reduce hallucinations in large vision-language models (LVLMs): object-aware visual grounding and reasoning-enhanced prompting. Both methods target different causes of hallucinations and were evaluated across object-centric and reasoning-heavy datasets to measure their effectiveness.

The visual grounding approach focused on enhancing factual alignment by integrating object detection outputs from YOLOv8 and DINO-X into the prompting pipeline of LLaVA. This method provided the LVLM with explicit, image-derived context, anchoring its responses to observable visual elements. Intuitively, LVLMs are often trained with images paired with global descriptions. However, training data typically lacks details like object positions, object attributes, and descriptions of non-salient objects (Chen et al., 2023), which prevents the model from aligning detailed object embeddings with contextual text embeddings. Using an object detector - trained specifically to recognize a wide range of objects in image - can introduce finer-grained precision into the language model, especially for details that are rarely mentioned in global descriptions.

Evaluations on the POPE adversarial dataset demonstrated measurable improvements in accuracy, recall, and F1 score. While DINO-X offered superior performance on rare and relational objects, YOLOv8 proved effective at large scale and yielded substantial gains using lightweight, real-time detection. These results confirm that enriching prompts with grounded object data can meaningfully reduce hallucinations in tasks requiring visual fidelity.

In parallel, we investigated language-based methods that restructure the model's reasoning process through Chain of Thought and Chain of Verification prompting. These techniques encouraged step-by-step reasoning, self-reflection, and explicit answer verification. Our comprehensive evaluation across POPE and Visual Puzzles datasets reveals that hallucination mitigation effectiveness is highly model-dependent and task-specific. These contrasting results suggest that hallucination mitigation strategies should be selectively applied based on both model architecture and task demands—enhancing weaker models on simple recognition tasks while potentially preserving direct inference for models with strong baseline capabilities or when facing complex reasoning challenges.

Across both strategies, our results highlight a key insight: hallucination reduction is highly task- and model-dependent. Visual grounding excels at reducing hallucinations related to object presence or misidentification, while reasoning-based prompting helps address logic and inference errors. Used together or in parallel, these approaches offer a more robust and generalizable framework for improving LVLM reliability. Future work may focus on hybrid models that fuse these methods more deeply—such as joint fine-tuning or multi-modal reasoning chains—to further advance hallucination mitigation in complex, real-world scenarios.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *Preprint*, arXiv:1906.02569.

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

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *Preprint*, arXiv:2311.16479.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. 2024. Real-time flying object detection with yolov8. *Preprint*, arXiv:2305.09972.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li, Kent Yu, and Lei Zhang. 2025. Dino-x: A unified vision model for open-world object detection and understanding. *Preprint*, arXiv:2411.14347.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Data-augmented phrase-level alignment for mitigating object hallucination. *arXiv preprint arXiv:2405.18654*.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. 2025. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating multimodal llm hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8460–8468.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2025. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, 133(2):825–843.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Appendix

A.1 Results for POPE Dataset evaluations

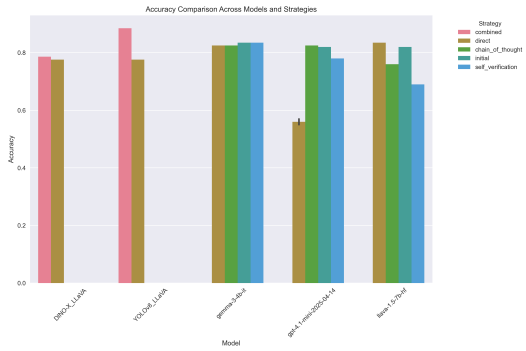

Figure 2: Accuracy comparison between strategies and models for POPE testing

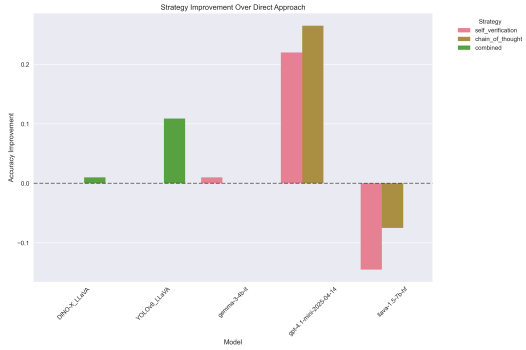

Figure 3: Strategy improvement vs baseline for POPE testing

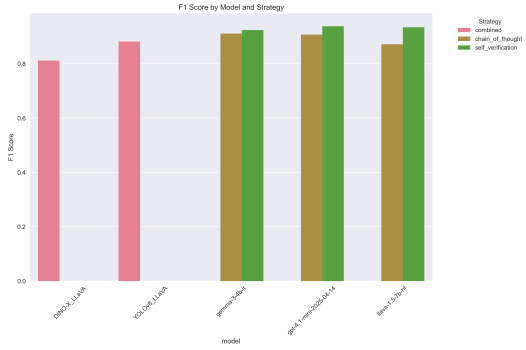

Figure 4: F1 Score for POPE testing

A.2 Results for VisualPuzzles Dataset evaluations

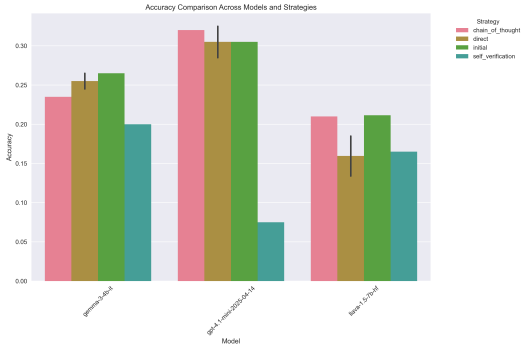

Figure 5: Accuracy comparison between strategies and models for VisualPuzzles testing

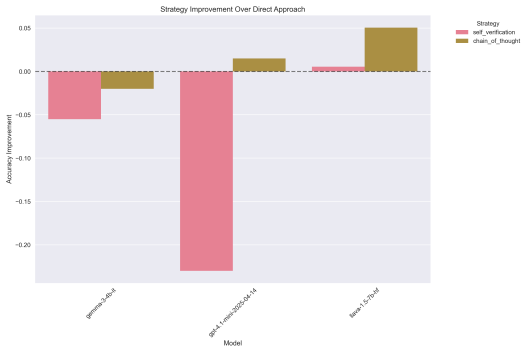

Figure 6: Strategy improvement vs baseline for VisualPuzzles testing

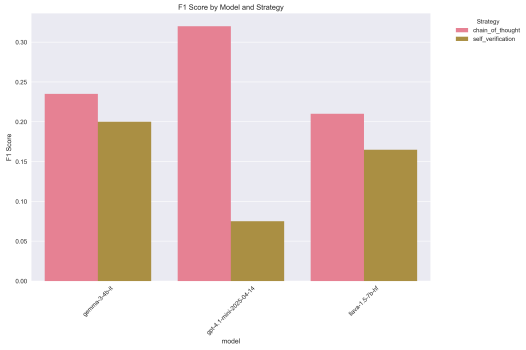

Figure 7: F1 Score for VisualPuzzles testing