
Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance

Linxi Zhao^{*1} Yihe Deng^{*2} Weitong Zhang³ Quanquan Gu²

Abstract

The advancement of Large Vision-Language Models (LVLMs) has increasingly highlighted the critical issue of their tendency to hallucinate non-existing objects in the images. To address this issue, previous works focused on using specially curated datasets or powerful LLMs to rectify the outputs of LVLMs. However, these approaches require either costly training or fine-tuning, or API access to proprietary LLMs for post-generation correction. In response to these limitations, we propose **Mitigating hallucinAtion via image-gRounded guIdaNcE** (MARINE), a framework that is both *training-free* and *API-free*. MARINE effectively and efficiently reduces object hallucinations during inference by introducing image-grounded guidance to LVLMs. This is achieved by leveraging open-source vision models to extract object-level information, thereby enhancing the precision of LVLM-generated content. Our framework's flexibility further allows for the integration of multiple vision models, enabling more reliable and robust object-level guidance. Through comprehensive evaluations across 5 popular LVLMs with diverse evaluation metrics and benchmarks, we demonstrate the effectiveness of MARINE, which even outperforms existing fine-tuning-based methods. Remarkably, it reduces hallucinations consistently in GPT-4V-assisted evaluation while maintaining the detailedness of LVLMs' generations. We release our code at <https://github.com/Linxi-ZHAO/MARINE>.

1 Introduction

The advent of Large Language Models (LLMs) has motivated advancements in extending their remarkable capabilities to multimodal data. Grounded in the development of pre-trained vision-language models (Radford et al., 2021; Jia et al., 2021; Alayrac et al., 2022) that align visual and textual embedding spaces, Large Vision Language Models (LVLMs) have gained substantial attention in both architectural development (Liu et al., 2023d; Zhu et al., 2023; Ye et al., 2023; Dai et al., 2023a; Gao et al., 2023), alignment (Yu et al., 2024; Zhou et al., 2024; Deng et al., 2024) and benchmarking datasets (Xu et al., 2023; Lu et al., 2024; Zhang et al., 2024a). However, similar to the hallucination issues in textual LLMs (Ji et al., 2023), where irrelevant content is generated with input prompts, LVLMs face a specific challenge known as object hallucination: generating non-existing objects for a given image (Li et al., 2023b; Wang et al., 2023b; Zhou et al., 2023; Fu et al., 2023; Lovenia et al., 2023; Jing et al., 2023). Such a problem is particularly concerning as it compromises the model's accuracy and reliability, especially considering the growing application of LVLMs to safety-critical downstream tasks such as medical imaging (Chambon et al., 2022; Bazi et al., 2023).

In response to the pressing issue of object hallucinations in LVLMs, early attempts (Liu et al., 2023a;b; Gunjal et al., 2023; Wang et al., 2023a) focused on addressing the bias by curating high-quality datasets for fine-tuning or leveraging advanced GPT queries (Yin et al., 2023), such as GPT-4, to post-process the generated captions. However, these methods can be infeasible to implement. For instance, creating extensive, high-quality datasets for fine-tuning LVLMs is costly and requires significant human annotation. Additionally, relying on advanced GPT models for post-processing is expensive and can raise privacy concerns, especially in sensitive fields like medical imaging. Most importantly, these approaches do not address the *intrinsic* causes of object hallucination in LVLMs.

In this paper, we investigate the intrinsic causes of object hallucination in LVLMs. Specifically, these deficiencies may stem from the three main components of the LVLMs: 1) insufficient visual context provided by the visual encoder (Zhang et al., 2023b), 2) distortion or loss of visual information during the projection from vision to text space,

^{*}Equal contribution . ¹Department of Computer Science, Cornell University, Ithaca, NY, USA ²Department of Computer Science, University of California, Los Angeles, CA, USA ³School of Data Science and Society, UNC, Chapel Hill, NC, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

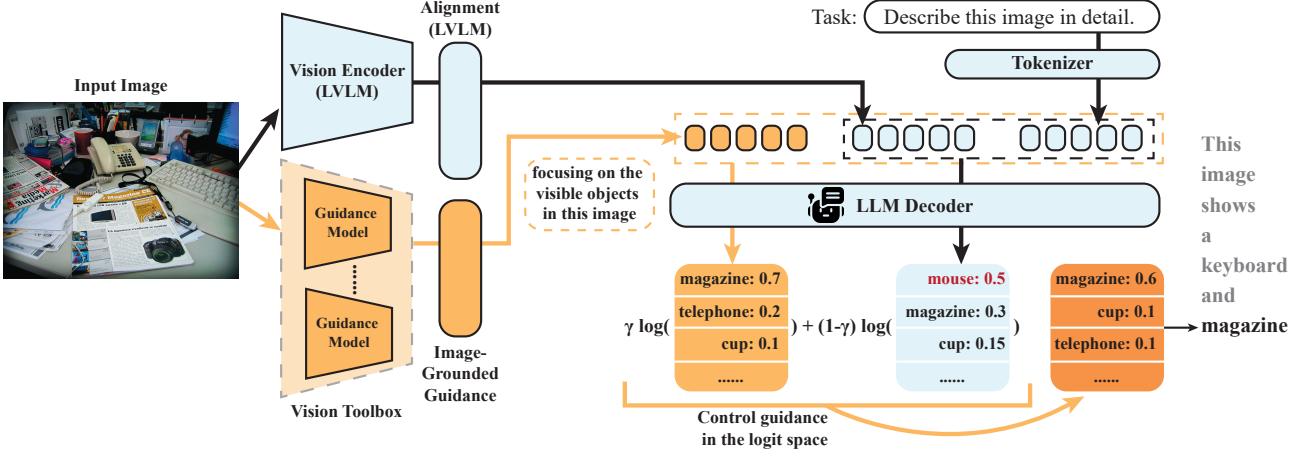


Figure 1. Illustration of MARINE framework, which introduces a vision toolbox with one or multiple guidance models to enrich the visual context of the original LVLM. The output logits are controlled to place more importance on the guided generation with the guidance strength γ .

and 3) inherent hallucinations common in general language models. To address the first two LVLM-specific causes, we introduce **Mitigating halluCinAtion via image-gRounded guIdaNcE** (MARINE). MARINE mitigates hallucination issues arising from the visual encoder and information distortion during cross-modal alignment by leveraging external guidance from image-grounded models, such as object detection models. Our approach leverages the inherent advantage of these image-grounded models, which are specifically designed and trained for more detailed visual information extraction. These models provide higher quality, fine-grained visual encoding compared to the standard visual encoders in LVLMs, which are primarily optimized for grasping the overall context of an image. Furthermore, we integrate the guidance from image-grounded models into text descriptions, allowing the LVLM to process the information without requiring additional alignment procedures. As a result, MARINE is a training-free, API-free method that addresses object hallucination at inference time by targeting its two root causes.

As shown in Figure 1, MARINE incorporates one or more image-grounding models to enrich the visual context of LVLMs. The guidance are then aggregated as prompt input to the LLM decoder to improve the response quality. Empirical evaluations are conducted on five widely-recognized LVLMs across benchmarks including MSCOCO (Lin et al., 2014), LLaVA-QA90 task (Liu et al., 2023d), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019). We present results based on guidance from a aggregated source of DETractor (DETR) (Carion et al., 2020) and RAM++ (Huang et al., 2023b). We also include ideal results based on ground truth object oracle, denoted as MARINE-Truth. Our experimental

results demonstrate that, in comparison with state-of-the-art algorithms, MARINE exhibits further reduced hallucination, as measured by popular hallucination metrics such as CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b), as well as GPT-4V’s evaluation. These results confirm that MARINE can effectively mitigate object hallucinations without requiring additional training resources or access to proprietary LLMs. To summarize, our contribution are listed as follows:

- We introduce MARINE, a universal framework and aggregating a toolbox of image-grounded visual models to guide the generation process of LVLMs. MARINE leverages the intrinsic advantages of these visual models in providing the detailed information of the input image and help mitigate the hallucinations in LVLMs.
- Through extensive evaluations on various datasets, we demonstrate that MARINE consistently outperform the baselines in hallucination mitigation while maintaining overall performance across multiple tasks (image captioning, VQA).
- MARINE provides a favorable trade-off between latency and accuracy, with the lowest computational overhead compared to existing baselines, which positions MARINE as a practical and scalable solution for real-world applications without significant computational cost.

2 Related Work

2.1 Object Hallucination in Large Vision-Language Models

The hallucination issue in Large Vision-Language Models (LVLMs) (Liu et al., 2023d; Zhu et al., 2023; Ye et al., 2023; Dai et al., 2023a; Gao et al., 2023) has drawn significant attention, as highlighted by studies (Li et al., 2023b; Wang

et al., 2023b; Zhou et al., 2023; Fu et al., 2023; Lovenia et al., 2023). Notably, different from textual LLMs, LVLMs are prone to a unique type of hallucination called ‘object hallucination’ (Rohrbach et al., 2018), where the model falsely perceives the presence of non-existent objects in images. Efforts to address this problem in LVLMs include fine-tuning approaches using vision-language datasets (Liu et al., 2023b; Gunjal et al., 2023), as well as GPT-assisted methods such as those by Zhai et al. (2023). Notably, Yin et al. (2023) proposed a training-free approach using GPT-3.5 for hallucination correction.

Concurrently, Leng et al. (2023) introduced Visual Contrastive Decoding (VCD), a technique that applies noise to image inputs and penalizes logit outputs of these corrupted images. Huang et al. (2023a) enhanced beam-search decoding with the Over-trust Penalty and Retrospection-Allocation Strategy (OPERA), which penalizes over-trust and refines token selection based on previous outputs. HALC (Chen et al., 2024) employs adaptive focal-contrast decoding to encourage LVLMs to focus on fine-grained visual information, while using a computationally intensive beam search algorithm. In addition, BRAVE (Kar et al., 2024) introduces a new architecture that combines features from multiple vision encoders. While not directly targeting hallucination, it shares the key insight of leveraging diverse visual signals to improve grounding.

2.2 Controllable Generation

Controllable text generation (Prabhumoye et al., 2020; Hu & Li, 2021; Zhang et al., 2023a) has emerged as a vital research domain, focusing on the generation of natural sentences with controllable attributes such as persona (Prabhumoye et al., 2020; Hu & Li, 2021; Zhang et al., 2023a) and politeness (Niu & Bansal, 2018; Madaan et al., 2020). Among the various approaches, fine-tuning has been recognized as the most straightforward approach, achieved either through full fine-tuning (Li & Liang, 2021; Ouyang et al., 2022; Carlsson et al., 2022) or integrating tunable adaptors (Lin et al., 2021; Ribeiro et al., 2021). While fine-tuning has been effective in a wide range of applications, it is also expensive in computation as the size of LLMs is growing tremendously. Recently, there has been a development on controllable generation with diffusion models (Li et al., 2022; Lin et al., 2023b), extending to controllable text-to-image generation (Yang et al., 2023). Particularly, the use of classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2021) has become prominent in refining the quality of generated outputs. Most recently, Sanchez et al. (2023) applied classifier-free guidance to language models in the *single-modal* setting to improve their performance at inference time. Our approach methodologically resembles classifier-free guidance for LVLMs’ text generation, while specifically addressing the *multi-modal* context and focusing on reducing hallucinations.

3 Preliminaries

Generative language models. Let p_{θ} denotes an LLM parameterized by θ . Consider a sequence $\mathbf{x} = [x_1, \dots, x_n]$ as the input prompt, where each x_i is a token from a predefined vocabulary. The LLM then generates the response sequence $\mathbf{y} = [y_1, \dots, y_m]$ by sampling from the conditional probability distribution $p_{\theta}(\cdot|\mathbf{x})$, where y_t denotes individual token for $1 \leq t \leq m$. The conditional distribution $p_{\theta}(\mathbf{y}|\mathbf{x})$ can therefore be expressed as $p_{\theta}(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^m p_{\theta}(y_t|\mathbf{x}, \mathbf{y}_{<t})$, where $\mathbf{y}_{<t} = [y_1, \dots, y_{t-1}]$ for $t > 1$ and is empty for $t = 1$. In the case of LVLMs, visual tokens $\mathbf{v} = [v_1, \dots, v_k]$ are additionally included. These tokens are generated from a pre-trained visual encoder and mapped into the token space through a linear projection. The conditional distribution of output \mathbf{y} given the visual tokens \mathbf{v} and textual prompt \mathbf{x} is expressed as $p_{\theta}(\mathbf{y}|\mathbf{v}, \mathbf{x}) = \prod_{t=1}^m p_{\theta}(y_t|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})$, where p_{θ} is approximated by LVLMs.

Guidance in generative models. The process of a guided generation involves getting the output \mathbf{y} conditioned on input \mathbf{x} , which encodes the desired properties of the output \mathbf{y} . This guidance can be generally added to the model by two distinct approaches: classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (Ho & Salimans, 2021). As a top-level view, both methods formulate the conditional probability distribution of output \mathbf{y} conditioned on guidance \mathbf{x} as

$$p(\mathbf{y}|\mathbf{x}) \propto p_{\theta}(\mathbf{y})p(\mathbf{x}|\mathbf{y})^{\gamma}, \quad (3.1)$$

where $p_{\theta}(\mathbf{y})$ is the original generative model and $p(\mathbf{x}|\mathbf{y})$ is the posterior distribution of \mathbf{x} given \mathbf{y} and γ is the guidance strength. In the classifier guidance, the posterior distribution $p(\mathbf{x}|\mathbf{y})$ in (3.1) is replaced by a classifier $p_{\phi}(\mathbf{x}|\mathbf{y})$ parameterized by ϕ , which requires additional training step and calculating $\nabla_{\mathbf{x}} \log p_{\phi}(\mathbf{x}|\mathbf{y})$. The classifier-free guidance, on the other hand, removes the necessity of the parameterized classifier f_{ϕ} . Instead, according to the Bayes rule, the posterior distribution can be approximated by $p_{\theta}(\mathbf{x}|\mathbf{y}) \propto p_{\theta}(\mathbf{y}|\mathbf{x})/p_{\theta}(\mathbf{y})$, where $p_{\theta}(\mathbf{y}|\mathbf{x})$ is the generative model when taking \mathbf{x} as prompt input. Plugging this back into (3.1) yields the guided distribution that can be approximated by

$$\widehat{p}_{\theta}(\mathbf{y}|\mathbf{x}) \propto p_{\theta}(\mathbf{y}) \cdot \frac{p_{\theta}(\mathbf{y}|\mathbf{x})^{\gamma}}{p_{\theta}(\mathbf{y})^{\gamma}} = \frac{p_{\theta}(\mathbf{y}|\mathbf{x})^{\gamma}}{p_{\theta}(\mathbf{y})^{\gamma-1}}.$$

As a result, the guided LLM \widehat{p}_{θ} places more importance on the prompt \mathbf{x} during generation with the increasing value of γ , thereby producing texts that better align with the desired behavior from the prompt (Sanchez et al., 2023).

4 Method

The existing architecture of LVLMs is composed of a visual encoder, a visual and textual domain alignment layer, and

the LLM itself. Therefore, besides the inherent language priors of LLMs (Biten et al., 2022), object hallucination may arise from (1) deficiencies in the visual encoder provide insufficient visual information (Zhang et al., 2023b) and (2) distortion or loss of visual information during the projection from vision to language space. To mitigate object hallucinations, we introduce MARINE, a framework containing two major components to address the previous challenges: (1) introducing additional visual information from a set of vision models and (2) using the additional aggregated visual features to guide the LVLM’s generation. In Figure 1, we present the framework overview.

4.1 Visual Guidance from Image-Grounded Features

To introduce image-grounded guidance to mitigate hallucinations, our approach integrates additional object detection models, which differ from the visual encoders used in LVLM that are usually pre-trained from CLIP (Radford et al., 2021). This integration leverages object detection models to extract detailed visual information from images. Upon acquiring extra visual information from different image-grounded models, we aggregate and translate the collected information into textual information. This aggregation can be done by the language model (Lin et al., 2023a) or rule-based algorithm (Bird et al., 2009). Such an information aggregation is effective and efficient, as it eliminates the necessity of fine-tuning the alignment layer while retaining the rich information encoded by various of image grounding models. We subsequently employ a simple prompt “focusing on the visible objects in this image:” and concatenate it with the aggregated object information, denoted as the guidance prompt \mathbf{c} .

4.2 Guided Text Generation with Visual Information

We tackle the object hallucination problem of LVLMs by placing importance on additional image-grounded information. In addition to the visual tokens \mathbf{v} extracted from the original LVLM and textual prompt \mathbf{x} , we extract the auxiliary visual tokens \mathbf{c} from the additional guidance models. The generation of the t -th token in the output \mathbf{y} of our classifier-free guided LVLM p_θ is expressed as

$$\hat{p}_\theta(y_t|\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) \propto \frac{p_\theta(y_t|\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t})^\gamma}{p_\theta(y_t|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})^{\gamma-1}},$$

where \mathbf{c} denotes our control guidance and γ is the control strength. The sampling of output generation is given by

$$\begin{aligned} \hat{p}_\theta(\mathbf{y}|\mathbf{v}, \mathbf{c}, \mathbf{x}) &= \prod_{t=1}^m \hat{p}_\theta(y_t|\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) \\ &\propto \prod_{t=1}^m \frac{p_\theta(y_t|\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t})^\gamma}{p_\theta(y_t|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})^{\gamma-1}} \\ &= \frac{p_\theta(\mathbf{y}|\mathbf{v}, \mathbf{c}, \mathbf{x})^\gamma}{p_\theta(\mathbf{y}|\mathbf{v}, \mathbf{x})^{\gamma-1}}. \end{aligned}$$

We can further view MARINE in the logit space, where the

t -th token is therefore sampled from the logit space by

$$\begin{aligned} \log \hat{p}_\theta(y_t|\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) &= \gamma \log p_\theta(\mathbf{y}|\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}) \\ &\quad + (1 - \gamma) \log p_\theta(\mathbf{y}|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}). \end{aligned}$$

This linear combination of logits implies that the conditional generation on the additional image-grounded guidance acts as a controllable gate. Only objects with relatively high probabilities in both branches could appear at top when sampling. Specifically, setting $\gamma = 0$ recovers the original LLM generation without control guidance and setting $\gamma = 1$ produces the LLM generation entirely based on the control. Meanwhile, for $\gamma \in (0, 1)$, MARINE yields a combination of the original generation $p_\theta(\mathbf{y}|\mathbf{v}, \mathbf{x})$ and the generation conditioned on the guidance $p_\theta(\mathbf{y}|\mathbf{v}, \mathbf{c}, \mathbf{x})$. This strikes a balance between a better ability to follow instructions to generate high-quality answers and the increased accuracy and detail in image descriptions. The formulation therefore shares resemblance to the classifier-free guidance introduced for LLMs (Sanchez et al., 2023), which places importance on the textual prompt itself to better align the LLM generation with user intention in the *single-modal* setting. We summarize MARINE in Algorithm 1. In detail, MARINE aggregates the collected visual information $\{\mathbf{c}_i\}_i$ using function Aggr., which can be a small language model for information aggregation (Lin et al., 2023a).

Algorithm 1 Mitigating halluCinAtion via image-gRounded guIdaNcE (MARINE)

- 1: **Input:** LLM parameter θ , input prompt \mathbf{x} , visual tokens \mathbf{v} from LVLM’s original vision tower
 - 2: **Input:** auxiliary visual tokens $\{\mathbf{c}_i\}_{i=1}^M$ from M image grounding models, guidance scale γ
 - 3: Initialize empty output $\mathbf{y} = []$.
 - 4: Aggregate visual information as textual prompt $\mathbf{c} = \text{Aggr.}(\{\mathbf{c}_i\}_{i=1}^M)$
 - 5: **for** $t = 0, 1, \dots, T$ **do**
 - 6: Construct unconditional input $\mathbf{x}_{\text{uncond}}^{(t)} = [\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}]$.
 - 7: Generate unconditional output logits using LLM: $\ell_{\text{uncond}}^{(t)} = \log p_\theta(\mathbf{x}_{\text{uncond}}^{(t)})$.
 - 8: Construct conditional input $\mathbf{x}_{\text{cond}}^{(t)} = [\mathbf{v}, \mathbf{c}, \mathbf{x}, \mathbf{y}_{<t}]$.
 - 9: Generate conditional output logits using LLM: $\ell_{\text{cond}}^{(t)} = \log p_\theta(\mathbf{x}_{\text{cond}}^{(t)})$.
 - 10: Update output logits $\ell^{(t)} = \gamma \ell_{\text{cond}}^{(t)} + (1 - \gamma) \ell_{\text{uncond}}^{(t)}$.
 - 11: Sample token y_t from logit space denoted by $\ell^{(t)}$.
 - 12: Let $\mathbf{y} = [\mathbf{y}, y_t]$.
 - 13: **end for**
 - 14: **Output:** \mathbf{y} .
-

5 Experiments

In this section, we evaluate MARINE in mitigating object hallucinations across various LVLMs, showing that it outperforms state-of-the-art methods on established metrics across different question formats.

5.1 Experiment Setup

Models. To demonstrate the broad applicability of our approach across different LVLM architectures, we apply and evaluate MARINE to widely-used models including *LLaVA* (Liu et al., 2023d), *LLaVA-v1.5* (Liu et al., 2023c), *MiniGPT-v2* (Chen et al., 2023), *mPLUG-Owl2* (Ye et al., 2023) and *InstructBLIP* (Liu et al., 2023c). To address the object hallucination problems in text generation, we incorporate the DEtection TRansformer (DETR) (Carion et al., 2020) and RAM++ (Huang et al., 2023b) as the additional vision models for guidance.

Guidance from multiple sources. Our framework’s compatibility with various vision models allows for the incorporation of multiple sources to enhance precision and robustness. By considering object-level information from DETR and RAM++ simultaneously, we generate guidance that reflects consensus across these models. This approach significantly improves the accuracy and reliability of the guidance provided to the LVLM.

Datasets and evaluations. In alignment with established evaluations from previous studies (Dai et al., 2023b; Yin et al., 2023), we assess our method using the following metrics:

- Caption Hallucination Assessment with Image Relevance (*CHAIR*) (Rohrbach et al., 2018). It involves prompting the LVLMs to generate a description for the input image, and then comparing this generation with ground truth objects present in the image. CHAIR quantifies hallucination both at instance level and sentence level, respectively defined as CHAIR_I and CHAIR_S :

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}$$

$$\text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

In addition to these metrics, we incorporate an instance-level Recall score in our evaluation to evaluate whether the descriptions accurately include the necessary visual content from the image:

$$\text{Recall} = \frac{|\{\text{non-hallucinated objects}\}|}{|\{\text{all existing objects}\}|}$$

- Polling-based Object Probing Evaluation (*POPE*) (Li et al., 2023b). POPE formulates a binary classification task by prompting LVLMs with questions such as “Is there a keyboard in this image?” to answer “yes” or “no”. We specifically focus on the adversarial setting, which is considered the most challenging setting. Results for the random and popular settings are detailed in Appendix C. We report the accuracy and F1 score of the LVLMs’ responses, and the proportion of “yes” answers.

- *GPT-4V-aided Evaluation* (Yin et al., 2023). The GPT-4V-aided evaluation compares the outputs of two LVLM assistants using GPT-4V as a judge. In this evaluation, we utilize the LLaVA-QA90 task (Liu et al., 2023d) (including conversations, visual perceptions, and complex reasoning tasks) and additionally consider the image captioning task.

Consistent with Li et al. (2023b), we randomly sampled a subset of 500 images from MSCOCO (Lin et al., 2014) dataset for CHAIR evaluation. For the POPE evaluation, we created 3000 questions across three datasets—500 images each from MSCOCO, A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019). For the GPT-4V-aided evaluation, we utilized 90 questions from the LLaVA-QA90 task and randomly selected 50 MSCOCO images for image captioning task.

Baselines. In addition to comparing with the performance of the original LVLM sampling method, we also consider the following popular methods for mitigating hallucinations.

- *Greedy-Decoding*, which adopts the greedy sampling strategy, by generating tokens with the highest posterior probability to address hallucinations arising from.
- *LURE* (Zhou et al., 2023), which identifies and masks potentially hallucinated words and fine-tune a MiniGPT4 model to rectify object hallucinations in the generated descriptions.
- *Woodpecker* (Yin et al., 2023), which leverages GPT-3.5 to correct hallucinations in LVLM generation with five steps toward the correction.
- *VCD* (Leng et al., 2023), which distorts the image inputs to impose penalties on logit outputs.
- *OPERA* (Huang et al., 2023a), which penalizes logits to mitigate over-trust in beam-search decoding and adjusts token selection.

Lastly, the performance of MARINE improves in correlation with the advancement of the control guidance extractor used. Consequently, to demonstrate the potential upper bound of MARINE’s performance, we consider a version utilizing a ground-truth oracle extractor, which we denote as MARINE-Truth. Further details on model architectures, datasets and evaluation metrics are deferred to Appendix A.

Hyperparameter setting. The hyperparameters for our method are fixed across tasks, with key settings including a guidance strength of 0.7, score threshold for DETR at 0.95, a detection threshold for RAM++ of 0.68, and a greedy sampling approach with a random seed of 242.

5.2 Results

Experimental results on object hallucination metrics (CHAIR and POPE) are presented in Table 1 and 2. Overall, MARINE achieves superior performances across different LVLM architectures and evaluation metrics.

Table 1. Evaluation with CHAIR score across multiple LVLM architectures comparing our method with several baselines. We report CHAIR_S , CHAIR_I and the recall score. The **bold** numbers indicate the best results among the methods evaluated and the underlined numbers represent the second-best results. We show MARINE-Truth as a reference performance of MARINE.

Method	LLaVA			LLaVA-v1.5			MiniGPTv2			mPLUG-Owl2			InstructBLIP			Average		
	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$	$C_S \downarrow$	$C_I \downarrow$	$R \uparrow$
Greedy	26.6	10.5	47.4	8.8	4.6	41.1	8.2	<u>4.2</u>	41.1	6.2	3.4	38.8	5.0	3.2	33.2	11.0	5.2	40.3
LURE	33.8	11.6	54.8	38.9	11.2	56.3	36.2	11.4	54.6	33.9	10.8	55.9	38.1	12.1	54.5	36.2	11.4	55.2
Woodpecker	19.5	<u>8.9</u>	44.3	8.5	4.5	38.4	<u>7.5</u>	4.5	37.0	8.0	4.3	37.5	8.0	6.2	32.6	10.3	5.7	38.0
VCD	28.1	11.0	46.6	<u>7.3</u>	<u>4.1</u>	40.8	6.8	3.9	38.2	<u>5.9</u>	3.4	37.7	<u>2.4</u>	<u>1.5</u>	33.7	<u>10.1</u>	<u>4.8</u>	39.4
OPERA	<u>22.4</u>	9.9	43.6	11.0	6.7	40.2	9.2	5.0	41.3	5.8	<u>3.2</u>	38.4	4.6	2.7	<u>38.0</u>	10.6	5.5	40.3
MARINE	17.8	7.2	<u>50.8</u>	6.2	3.0	44.3	11.8	4.9	<u>49.7</u>	4.2	2.3	41.4	2.2	1.3	36.3	8.4	3.7	<u>44.5</u>
MARINE-Truth	19.6	5.1	79.0	6.0	2.5	55.3	12.6	3.8	70.5	3.8	1.7	48.0	3.0	1.8	35.9	8.9	2.9	57.5

Table 2. Evaluation with POPE score in adversarial setting across multiple LVLM architectures comparing our method with several baselines. We report the POPE accuracy (%), F1 score (%) and the yes ratio (%). The ideal yes ratio for a non-biased LVLM is 50%. The **bold** numbers indicate the best results among the methods evaluated and the underlined numbers represent the second-best results. We show MARINE-Truth as a reference performance of MARINE.

Method	LLaVA			LLaVA-v1.5			MiniGPTv2			mPLUG-Owl2			InstructBLIP			Average		
	Acc \uparrow	F1 \uparrow	Yes															
Greedy	51.8	67.4	97.7	79.4	<u>81.6</u>	61.6	82.7	81.7	<u>44.5</u>	72.5	77.5	72.4	<u>79.8</u>	81.4	58.6	73.2	77.9	67.0
LURE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Woodpecker	77.5	77.6	50.5	<u>80.5</u>	80.6	50.5	79.5	77.8	42.5	77.5	76.9	<u>47.5</u>	79.0	78.6	48.0	<u>78.8</u>	<u>78.3</u>	<u>47.8</u>
VCD	54.6	68.5	94.0	78.2	80.7	62.8	81.4	80.2	44.1	72.3	77.0	70.5	79.7	<u>80.9</u>	<u>56.7</u>	73.2	77.5	65.6
OPERA	51.7	67.4	98.0	77.5	80.1	63.2	<u>82.9</u>	<u>81.9</u>	44.3	70.3	<u>79.1</u>	84.6	<u>79.8</u>	81.4	58.6	72.4	78.0	69.7
MARINE	<u>66.9</u>	<u>72.9</u>	<u>72.3</u>	85.0	84.3	<u>45.7</u>	83.0	82.9	49.4	82.8	82.7	49.2	81.7	79.4	38.8	79.9	80.4	51.1
MARINE-Truth	75.6	80.1	72.3	92.0	92.5	57.0	86.9	88.3	62.5	93.4	93.8	56.2	93.8	93.8	51.0	88.3	89.7	59.8

Table 3. Results of GPT-4V-aided evaluation. The accuracy and detailedness metrics are on a scale of 10, and a higher score indicates better performance. The symbols \times and \checkmark indicate performance metrics without and with our method, respectively.

Task	Metrics	LLaVA		mPLUG-Owl2	
		\times	\checkmark	\times	\checkmark
LLaVA-QA90	Acc \uparrow	$5.82_{\pm 0.10}$	5.94 $_{\pm 0.05}$	$6.03_{\pm 0.13}$	6.35 $_{\pm 0.21}$
	Detail \uparrow	$4.59_{\pm 0.08}$	4.59 $_{\pm 0.08}$	$5.06_{\pm 0.05}$	5.16 $_{\pm 0.10}$
Image Captioning	Acc \uparrow	$5.27_{\pm 0.20}$	6.11 $_{\pm 0.23}$	$7.97_{\pm 0.25}$	8.63 $_{\pm 0.20}$
	Detail \uparrow	4.39 $_{\pm 0.29}$	$4.36_{\pm 0.17}$	$5.74_{\pm 0.24}$	6.19 $_{\pm 0.23}$

Results on CHAIR. CHAIR is a widely adopted benchmark for evaluating caption hallucination in LVLMs, comparing generated descriptions with ground-truth object annotations. It captures object-level precision through CHAIR_I (instance-level) and CHAIR_S (sentence-level), and we further report Recall to assess content coverage.

Table 1 shows that MARINE consistently outperforms existing approaches on all major metrics. It achieves the lowest average CHAIR_I and CHAIR_S scores and ranks second in Recall, reducing hallucination without sacrificing coverage. Compared to the second-best method, MARINE improves CHAIR_S by 1.7 points and CHAIR_I by 1.1 on average. The gains are particularly strong on LLaVA models, where hallucination drops by up to 8.8 points. In contrast, methods

such as LURE and Woodpecker are less effective across model variants.

Importantly, MARINE achieves performance comparable to MARINE -Truth, a variant that uses ground-truth object labels as guidance. This finding suggests that aggregating signals from multiple visual models offers a compelling alternative to manual supervision to reduce hallucination.

Results on POPE. POPE is designed to assess object-level grounding in LVLMs by testing their ability to answer yes/no questions about visual content. We focus on the adversarial setting, which presents challenging negatives and helps expose hallucination and biased answering tendencies. In Table 2, MARINE consistently outperforms all baselines, with average improvements of 6.7% in accuracy and 3.5% in F1 score over the original model outputs. Compared to the second-best method, Woodpecker, MARINE still maintains a 1.1% gain in accuracy and a 2.1% gain in F1. Beyond accuracy, MARINE also reduces the overconfident bias often seen in LVLMs' outputs. This is reflected in a more balanced “yes” ratio (closer to 50%, reflecting a 15.9% shift towards unbiased answers). This shift suggests that MARINE produces more trustworthy predictions by reducing the tendency toward overconfident affirmative responses.

Table 4. POPE results across three datasets. We report the average score under random, popular, adversarial settings. The detailed POPE results can be found in the appendix C. The **bold** numbers indicate the best results. The ideal yes ratio for a non-biased LVLM is 50%.

Dataset	w/MARINE	LLaVA			mPLUG-Owl2		
		Accuracy ↑	F1 ↑	Yes(%)	Accuracy ↑	F1 ↑	Yes(%)
MSCOCO	✗	54.2	68.5	95.5	76.7	80.4	68.2
	✓	72.2	76.4	66.9	85.5	85.0	46.5
A-OKVQA	✗	51.8	67.5	97.9	69.6	76.5	78.5
	✓	64.3	72.8	80.2	82.0	83.5	57.2
GQA	✗	52.0	67.6	97.8	73.7	78.7	72.6
	✓	62.5	71.8	81.8	80.1	80.6	51.1

Results on GPT-4V-aided evaluation. Following Yin et al. (2023), this GPT-4V-assisted evaluation provides a qualitative perspective that complements the numerical metrics of CHAIR and POPE, offering a more comprehensive assessment of model performance. As shown in Table 3, GPT-4V consistently assigns higher accuracy with equal detailedness scores to models enhanced by MARINE, highlighting its ability to produce more precise and detailed descriptions, which demonstrates the robustness of our method in real-world visual tasks. The evaluation prompt is detailed in Appendix A.5.

Additional results on other vision-language tasks. To further evaluate the generalizability of our approach beyond object hallucination and the MSCOCO dataset, we extended our evaluations to additional datasets including A-OKVQA and GQA and included more general caption quality metrics. As shown in Table 4, the POPE results demonstrate that our method consistently mitigates hallucinations across various datasets with different image distributions. Figure 2 presents a comprehensive evaluation of the image captioning task on MSCOCO and LLaVA-QA90, a comprehensive VQA dataset, using metrics including BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). These results demonstrate that, although our method primarily targets hallucination mitigation, it maintains the overall performance of LVLMs on broader tasks, with no significant trade-offs in caption or VQA quality.

Latency analysis Many existing approaches to mitigating object hallucination rely on post-generation correction models (Zhou et al., 2023; Zhai et al., 2023; Yin et al., 2023), external object detectors (Yin et al., 2023), or complex decoding strategies (Huang et al., 2023a; Leng et al., 2023), all of which introduce substantial computational overhead. To assess the practical efficiency of MARINE, we evaluate its latency compared to existing baselines on LLaVA-7B, as shown in Table 5.

Our measurements include the time required for additional forward passes through external vision models. These models contribute only marginal latency relative to the cost of autoregressive decoding in LVLMs. In general, MARINE

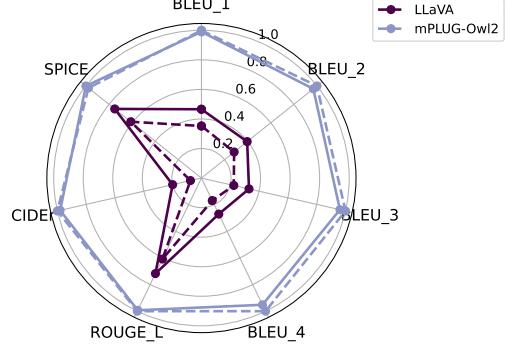


Figure 2. MARINE maintains or improves overall text quality on general metrics. Solid lines indicate models with MARINE, while dashed lines indicate the original models. Higher scores indicate better textual similarity to the reference outputs.

increases decoding time by just $1.98 \times$, the lowest among all baselines. This demonstrates that MARINE achieves the most favorable trade-off between latency and accuracy, which makes it suitable for real-world use. Detailed settings are provided in Appendix A.6.

5.3 Ablation Study

Why incorporate multiple image-grounded models? Different image-grounded models excel at capturing different aspects of visual information—some detect objects precisely, while others offer broader, fine-grained context. To understand whether combining these complementary signals leads to better guidance, we conduct an ablation comparing DETR and RAM++ individually versus in combination (Table 6). All variants are evaluated under the same decoding setup to ensure a fair comparison.

DETR allows for highly accurate object detection, while RAM++ excels in extensive recognition tasks, contributing fine-grained visual concepts. Their combination yields consistent improvements on CHAIR metrics, suggesting that aggregating multiple visual perspectives is important for effective hallucination mitigation.

What is the best way to integrate guidance from multiple models? When aggregating the outputs from multiple image-grounding models, the combination method can significantly affect guidance quality. We compare two strategies: taking the intersection or the union of detected objects. As shown in Table 7, the intersection-based approach consistently outperforms the union, significantly reducing hallucination. This suggests that enforcing agreement across models leads to more precise and trustworthy guidance, while union-based aggregation may introduce noisy or spurious information. The detailed experimental setup and prompt templates are provided in Appendix A.

How does control strength affect generation? To understand the impact of guidance strength in our decoding setup,

Table 5. Inference latency comparison. We report both the latency and the ratio to the latency of greedy decoding of the original LVLM model.

	Greedy	LURE	Woodpecker*	VCD	OPERA	MARINE (ours)
Training Cost	0	10min on A100 80G	0	0	0	0
Inference Latency ^(ms/token)	26.3 ($\times 1.0$)	179.9 ($\times 6.84$)	94.5 ($\times 3.59$)*	53.4 ($\times 2.03$)	185.1 ($\times 7.0$)	52.2 ($\times 1.98$)

* Woodpecker requires GPT API key access and the latency may depend on OPENAI API.

Table 6. Ablation study comparing the performance of combining DETR and RAM++ models versus using individual vision models. This approach leverages multiple object detectors to provide more reliable and robust object-level guidance, resulting in superior performance on CHAIR metrics.

Model	LLaVA		LLaVA-v1.5		mPLUG-Owl2	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Greedy	26.6	10.5	8.8	4.6	6.2	3.4
<i>Ensembling Models</i>						
MARINE	17.8	7.2	6.2	3.0	4.2	2.3
<i>Single Models</i>						
MARINE-DETR only	27.6	8.4	10.5	4.3	5.3	2.7
MARINE-RAM only	29.0	9.1	6.6	3.7	5.2	2.8

Table 7. Effect of Integration Methods for Image-Grounding Models.

Model	LLaVA		LLaVA-v1.5		mPLUG-Owl2	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Greedy	26.6	10.5	8.8	4.6	6.2	3.4
MARINE-intersection (ours)	17.8	7.2	6.2	3.0	4.2	2.3
MARINE-union	30.4	9.7	5.4	2.7	4.8	2.7

we vary the control weight γ , which balances the influence between the original LVLM generation and the generation conditioned on external image-grounded guidance.

Figure 3 shows that increasing guidance strength from 0 to 1 leads to a notable decrease in CHAIR scores. This trend suggests that higher guidance strength makes LVLMs rely more on image-grounded features, thereby enhancing their ability to produce accurate descriptions. It’s crucial to note that, although some models exhibit optimal performance at a guidance strength of $\gamma = 1$, excessively strong guidance can adversely affect the models’ ability to adhere to provided instructions. Experimental evidence is detailed in Appendix B.5. This observation highlights the necessity of having a balanced guidance strength that ensures high-quality, accurate outputs while adhering closely to the given instructions. Based on our findings, we recommend a guidance strength within the range of $\gamma \in (0.3, 0.7)$ as the most effective for maintaining this balance.

6 Conclusions, Limitations and Future Work

In this paper, we introduced a training-free and API-free framework MARINE to mitigate object hallucination in LVLMs during its text generation process. Leveraging a pre-trained object grounding vision encoder for a novel guid-

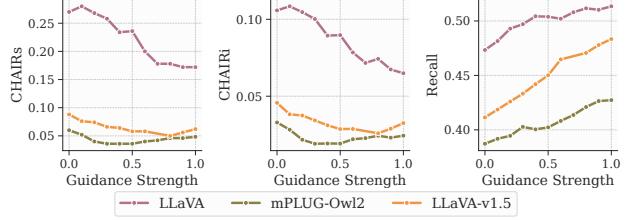


Figure 3. Ablation study on the effect of guidance strength (γ) on the performance of LLaVA, LLaVA-v1.5 and mPLUG-Owl2 using CHAIR metrics, with γ ranging from 0 to 1.

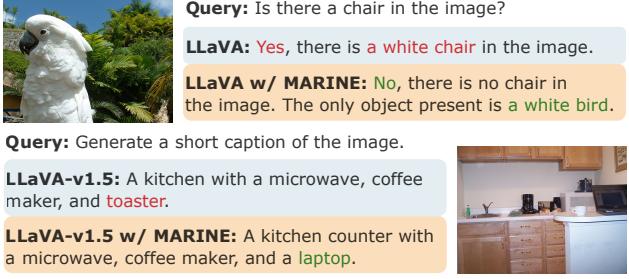


Figure 4. Hallucination mitigation examples by our proposed MARINE across multiple tasks. Hallucinated objects generated by the LVLM are highlighted in red.

ance framework in the multi-modal setting, MARINE effectively and cost-efficiently reduces the hallucinations of five widely-used LVLMs, as assessed by various metrics across different tasks. The inherent compatibility of the MARINE with various vision models and projection functions further underscores its flexibility. In contrast to post-generation correction methods, MARINE strikes a balance between efficiency, instruction-following ability and effectiveness in reducing object hallucinations.

Limitations and future work. While MARINE has demonstrated impressive performance by utilizing guidance from image-grounded models, there remains potential for further improvement through the integration of advanced aggregation methods, such as multi-agent debate (Du et al., 2023), into the MARINE framework. Additionally, although MARINE is specifically designed to mitigate object hallucination, which is the most significant issue in LVLMs, extending its application to address other types of hallucinations in both LLMs and LVLMs across a broader range of benchmarks would be highly advantageous.

Acknowledgments

We thank anonymous reviewers for their helpful comments. Part of this work was done while WZ was a PhD student at UCLA. WZ and QG are supported in part by NSF grants DMS-2323113, CPS-2312094, IIS-2403400, and the research fund from the UCLA-Amazon Science Hub. WZ was also supported by the UCLA dissertation year fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

Impact Statement

This paper introduces research aimed at advancing the field of Large Language Models. We are confident that our work will contribute to significant social benefits, particularly by enhancing the accountability of LLMs through the reduction of hallucinatory outputs. Our proposed method, MARINE, holds the potential to improve the fairness of LLM interactions by effectively reducing biased hallucinations. By mitigating hallucinations, MARINE has the potential to offer a positive social impact by ensuring that LVLMs generate more accountable responses. Despite this merit, MARINE cannot address prejudicial biases inherent in LLM prior knowledge, which could be a focus of future work. To the best of our knowledge, we have not identified any negative effects associated with our research that merit highlighting in this discussion.

References

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation, 2016.
- Bazi, Y., Rahhal, M. M. A., Bashmal, L., and Zuair, M. Vision-language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380, 2023.
- Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit.* ”O’Reilly Media, Inc.”, 2009.
- Biten, A. F., Gómez, L., and Karatzas, D. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1381–1390, 2022.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.
- Carlsson, F., Öhman, J., Liu, F., Verlinden, S., Nivre, J., and Sahlgren, M. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6837–6857, 2022.
- Chambon, P., Bluethgen, C., Langlotz, C. P., and Chaudhari, A. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.
- Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., and Elhoseiny, M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.
- Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., and Zhou, J. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., and Wang, S. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation, 2024. URL <https://arxiv.org/abs/2310.18235>.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023a.
- Dai, W., Liu, Z., Ji, Z., Su, D., and Fung, P. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2128–2140, 2023b.
- Deng, Y., Lu, P., Yin, F., Hu, Z., Shen, S., Zou, J., Chang, K.-W., and Wang, W. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*, 2024.
- Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.
- Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023.
- Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., and Qiao, Y. Llama-adapter v2: Parameter-efficient visual instruction model, 2023.
- Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.
- Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
- Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023. URL <https://arxiv.org/abs/2303.11897>.
- Hu, Z. and Li, L. E. A causal lens for controllable text generation. *Advances in Neural Information Processing Systems*, 34:24941–24955, 2021.
- Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., and Yu, N. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023a.
- Huang, X., Huang, Y.-J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., and Zhang, L. Open-set image tagging with multi-grained text supervision, 2023b. URL <https://arxiv.org/abs/2310.15200>.
- Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.
- Jing, L., Li, R., Chen, Y., Jia, M., and Du, X. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023.
- Kar, O. F., Tonioni, A., Poklukar, P., Kulshrestha, A., Zamir, A., and Tombari, F. Brave: Broadening the visual encoding of vision-language models, 2024. URL <https://arxiv.org/abs/2404.07204>.
- Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.
- Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.
- Li, X., Thickstun, J., Gulrajani, I., Liang, P. S., and Hashimoto, T. B. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022.
- Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.
- Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.

- Lin, S.-C., Li, M., and Lin, J. Aggretreiver: A simple approach to aggregate textual representations for robust dense passage retrieval. *Transactions of the Association for Computational Linguistics*, 11:436–452, 2023a.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.
- Lin, Z., Madotto, A., Bang, Y., and Fung, P. The adapter-bot: All-in-one controllable conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 16081–16083, 2021.
- Lin, Z., Gong, Y., Shen, Y., Wu, T., Fan, Z., Lin, C., Duan, N., and Chen, W. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning*, pp. 21051–21064. PMLR, 2023b.
- Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., and Ramanan, D. Evaluating text-to-visual generation with image-to-text generation, 2024. URL <https://arxiv.org/abs/2404.01291>.
- Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.
- Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2023b.
- Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023c.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023d.
- Liu, S., Ye, H., Xing, L., and Zou, J. Reducing hallucinations in vision-language models via latent space steering, 2024a. URL <https://arxiv.org/abs/2410.15778>.
- Liu, S., Zheng, K., and Chen, W. Paying more attention to image: A training-free method for alleviating hallucination in lvmms. *arXiv preprint arXiv:2407.21771*, 2024b.
- Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., and Fung, P. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*, 2023.
- Lu, J., Yang, J., Batra, D., and Parikh, D. Neural baby talk. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228, 2018. doi: 10.1109/CVPR.2018.00754.
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Math-vista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024.
- Madaan, A., Setlur, A., Parekh, T., Poczós, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., and Prabhumoye, S. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1869–1881, 2020.
- Niu, T. and Bansal, M. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389, 2018.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
- Petryk, S., Chan, D., Kachinthaya, A., Zou, H., Canny, J., Gonzalez, J., and Darrell, T. ALOHa: A new measure for hallucination in captioning models. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nacl-short.30. URL <https://aclanthology.org/2024.nacl-short.30/>.
- Prabhumoye, S., Black, A. W., and Salakhutdinov, R. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1–14, 2020.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

- Ribeiro, L. F., Zhang, Y., and Gurevych, I. Structural adapters in pretrained language models for amr-to-text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4269–4282, 2021.
- Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.
- Sanchez, G., Fan, H., Spangher, A., Levi, E., Ammanamanchi, P. S., and Biderman, S. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*, 2023.
- Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Vedantam, R., Zitnick, C. L., and Parikh, D. Cider: Consensus-based image description evaluation, 2015.
- Wan, D., Cho, J., Stengel-Eskin, E., and Bansal, M. Contrastive region guidance: Improving grounding in vision-language models without training. *arXiv preprint arXiv:2403.02325*, 2024.
- Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023a.
- Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b.
- Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., and Luo, P. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.
- Yang, L., Zheng, Z., Chen, B., Zhao, Z., Lin, C., and Shen, C. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection, 2025. URL <https://arxiv.org/abs/2412.13817>.
- Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023.
- Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., and Chen, E. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.
- Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. 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*, pp. 13807–13816, 2024.
- Zhai, B., Yang, S., Xu, C., Shen, S., Keutzer, K., and Li, M. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pp. arXiv–2310, 2023.
- Zhang, H., Song, H., Li, S., Zhou, M., and Song, D. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37, 2023a.
- Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Gao, P., et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*, 2024a.
- Zhang, Y., Qian, S., Peng, B., Liu, S., and Jia, J. Prompt highlighter: Interactive control for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13215–13224, 2024b.
- Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models, 2023b.
- Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.
- Zhou, Y., Cui, C., Rafailov, R., Finn, C., and Yao, H. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M.
Minigpt-4: Enhancing vision-language understanding
with advanced large language models. *arXiv preprint*
arXiv:2304.10592, 2023.

A Experiment Setup

We conduct all of the experiments using 8 A6000 GPU with 48GB GPU memory. Each single experiment can be run on a single A6000 GPU.

A.1 Model Architectures

In Table 8, we provide detailed descriptions of the LVLM architectures used in our experiments. These LVLMs respectively leverage the pre-trained vision encoder of the models we listed, which are all based on the Vision Transformer (ViT) (Dosovitskiy et al., 2020) architecture.

Table 8. Details of the LVLM architectures that we used in our paper.

Model	Vision encoder	LLM
LLaVA (Liu et al., 2023d)	CLIP-L (Radford et al., 2021)	LLaMA-2-7B-Chat (Touvron et al., 2023)
LLaVA-v1.5 (Liu et al., 2023c)	CLIP-L-336px (Radford et al., 2021)	Vicuna-v1.5-7B (Chiang et al., 2023)
MiniGPT-v2 (Chen et al., 2023)	EVA-G (Fang et al., 2023)	LLaMA-2-7B-Chat (Touvron et al., 2023)
mPLUG-OWL2 (Ye et al., 2023)	CLIP-L (Radford et al., 2021)	LLaMA-2-7B (Touvron et al., 2023)
InstructBLIP (Dai et al., 2023a)	BLIP-2 (Li et al., 2023a)	Vicuna-v1.1-7B (Chiang et al., 2023)

A.2 Descriptions about Additional Metrics

In Figure 2, we evaluate the text quality of the outputs generated with MARINE using general metrics as follows:

- *BLEU* (Papineni et al., 2002) measures how well the generated translation matches the reference translations in terms of n-gram overlap.
- *ROUGE-L* (Lin, 2004) measures the quality of a machine-generated summary by comparing it to one or more reference summaries.
- *CIDEr* (Vedantam et al., 2015) assesses the quality of image captioning models. It focuses on evaluating how well the generated captions align with human consensus.
- *SPICE* (Anderson et al., 2016) focuses on assessing the semantic similarity between the generated captions and reference captions.

A.3 Prompt Templates

For each query, we randomly select a prompt template from the available template list, as shown in Table 9.

A.4 Details of Baselines

Specifically, the hyperparameters for LURE (Zhou et al., 2023), VCD (Leng et al., 2023), OPERA (Huang et al., 2023a) are reported in Table 10, 11 and 12 respectively. We strictly followed the original implementations and default hyperparameters described in their papers to reproduce the results for each baseline.

A.5 Experiment Setting for Hallucination Evaluations

Key factors that potentially affect the hallucination evaluation outcomes, including the evaluation dataset and prompt template, LVLM’s sampling strategy and batched generation techniques, and guidance strength, are detailed in this section. The hyper-parameters setting for MARINE and overall experiment settings are shown in Table 13 and 14.

Experiment setting for CHAIR evaluation. We adopt the same prompt “Generate a short caption of the image.” as utilized by Li et al. (2023b). The hyperparameters are fixed, including a guidance strength of 0.7, score threshold for DETR at 0.95, a detection threshold for RAM++ of 0.68, a maximum token length of 64, and a greedy sampling approach with a random seed of 242.

For the calculation of CHAIR metrics, we referenced the 80 object categories annotated in the MSCOCO dataset, following Rohrbach et al. (2018). Besides, we employed the synonym list from Lu et al. (2018) to align synonymous words in the generated text with MSCOCO object categories. Additionally, due to the cost considerations associated with the GPT-3.5 API, we limited our analysis to 200 samples for Woodpecker correction for each model and reported the result in Table 1.

Experiment setting for POPE evaluation. POPE is a flexible approach to evaluating hallucinations in LVLMs, which formulates a binary classification task by prompting LVLMs with questions such as “Is there a keyboard in this image?” to answer “yes” or “no”. Following Li et al. (2023b), we created 3000 POPE questions across three datasets—500 images each from MSCOCO, A-OKVQA, and GQA for the POPE evaluation. We reported the adversarial settings in Table 2, the most challenging setting, which constructs POPE questions from the top-k most frequently co-occurring but absent objects. Additionally, in Table 4, we reported the average scores under random, popular, adversarial settings across MSCOCO,

Table 9. Details of the LVLM architectures that we used in our paper.

Template Type	Prompt Template
MARINE-intersec	<p>This image contains <OBJECT_GROUNDING>. Based on this, <QUERY></p> <p>The image contains the following objects: <OBJECT_GROUNDING>. Given these detected objects, <QUERY></p> <p>This image shows the following objects: <OBJECT_GROUNDING>. Using this information, <QUERY></p> <p>The objects found in this image are: <OBJECT_GROUNDING>. Considering this list of objects, <QUERY></p>
POPE task	<p>This image contains only the following objects: <OBJECT_GROUNDING>. Do not assume anything beyond these objects. Based solely on this list, <QUERY></p> <p>The detected objects in the image are: <OBJECT_GROUNDING>. Answer based only on these objects. <QUERY></p> <p>This image shows the following objects: <OBJECT_GROUNDING>. You must answer using only the objects in this list. Given these detected objects, <QUERY></p> <p>The objects found in this image are limited to: <OBJECT_GROUNDING>. You should rely strictly on this list of objects and make no other guesses. Based on this, <QUERY></p>
MARINE-union	<p>List of detected objects in the image:</p> <p><OBJECT_GROUNDING_A></p> <p><OBJECT_GROUNDING_B></p> <p>Based on the detected objects above, <QUERY></p> <p>The most prominent objects detected are:</p> <p><OBJECT_GROUNDING_A></p> <p><OBJECT_GROUNDING_B></p> <p>Given these findings, <QUERY></p> <p>The following objects were detected in the image:</p> <p><OBJECT_GROUNDING_A></p> <p><OBJECT_GROUNDING_B></p> <p>With this information, <QUERY></p> <p>Here is a list of all objects detected in the image:</p> <p><OBJECT_GROUNDING_A></p> <p><OBJECT_GROUNDING_B></p> <p>Do not infer or hallucinate any additional objects. Using only the detected objects, <QUERY></p>

Table 10. LURE (Zhou et al., 2023) Hyperparameter Settings

Parameters	Value
Uncertainty Threshold γ	0.9
Position Threshold ι	0.8

Table 11. VCD (Leng et al., 2023) Hyperparameter Settings

Parameters	Value
Amplification Factor α	1
Adaptive Plausibility Threshold	0.1
Diffusion Noise Step	500

Table 12. OPERA (Huang et al., 2023a) Hyperparameter Settings

Parameters	Value
Self-attention Weights Scale Factor θ	50
Attending Retrospection Threshold	25
Beam Size	5
Attention Candidates	1
Penalty Weights	1

Table 13. MARINE Hyperparameter Settings. The settings are fixed depending on the question-answering tasks.

Parameters	Value
<i>Guidance</i>	
Guidance Strength	0.7
score threshold for DETR	0.95
Detect Threshold for RAM++	0.68
<i>Generation</i>	
Max Token Length	64
Sampling	Greedy
Random Seed	242

Table 14. Batch size for LVLM generation is fixed across all experiments unless otherwise noted. To expedite the evaluation process, we employed the batched generation. We avoid the negative impact of batched generation by adopting left padding if the LVLM does not explicitly assign the padding strategy for inference.

Model	LLaVA	LLaVA-v1.5	MiniGPTv	mPLUG-Owl2	InstructBLIP
Batch Size	16	4	32	16	16

A-OKVQA, and GQA datasets. The full POPE results are in Tabel 16.

Similarly, we constrained our analysis to 200 samples for Woodpecker correction for each model due to the high costs associated with the GPT API. The outcomes of this analysis are detailed in Table 2.

Experiment setting for GPT-4V-aided evaluation. The GPT-4V-aided evaluation compares the outputs of two LVLM assistants using GPT-4V as a judge. We prompted GPT-4V to assess the quality of the generated outputs, scoring them out of 10 in two aspects:

- *Accuracy*: how accurately each assistant describes the image;
- *Detailedness*: the richness of necessary details in the response.

As shown in Table 15, the assessment prompt template we used is slightly different from that of Yin et al. (2023). Specifically, we include the original question for a task-orientated evaluation and exclude prompts that describe Woodpecker-specific output formats like object bounding boxes.

Experiment setting for ablation study. To explore different methods of integrating image-grounding models, we investigate the intersection and union of detected objects, with integration based on synonyms using the NLTK package. To quantitatively assess the influence of guidance strength, we varied it from 0 to 1, as shown in Figure 7. These quantitative experiments were conducted using the same setting as those in CHAIR evaluation. For qualitative analysis, we selected guidance strength from a recommended range of $\gamma \in (0.3, 0.7)$.

A.6 Experiment Setting on Other Vision-Language Tasks

Experiment setting for text quality analysis. For text quality analysis, we adopted 90 visual questions from the LLaVA-QA90¹ task (including conversations, visual perceptions, and complex reasoning subtasks), and randomly selected 500 MSCOCO images for image captioning task. Following Liu et al. (2023d), we adpted the response generated by text-only GPT-4 (0314) with the context captions/boxes provided. answers given by GPT-4 as references for LLaVA-QA90 task and used image captions provided in MSCOCO annotations as references for image captioning task.

In Table 17 and Table 18, we present a detailed evaluation on the image captioning task for both MSCOCO and LLaVA-QA90 using metrics including BLEU, ROUGE, CIDEr and SPICE. The corresponding figure result is shown in Figure 2.

Experiment setting for latency analysis. We compared our method with existing baselines in terms of the trade-off between inference cost and the effectiveness of reducing object hallucinations, as shown in Table 5. For post-correction baselines such as Woodpecker and LURE, we first prompted LLaVA (llava-llama-2-7b-chat-lightning-preview) to generate captions and then measure the latency of generating the corrected outputs. The total latency for post-correction baselines includes both the generation and correction processes. For decoding methods such as VCD, OPERA and our method, we measured the latency of LLaVA generating captions directly.

We prompted the models with “Generate a short caption of the image.” on 500 MSCOCO images with a batch size of 1 and

¹https://github.com/haotian-liu/LLaVA/blob/main/playground/data/coco2014_val_gpt4_qa_30x3.jsonl

Table 15. Prompt template for GPT-4V-aided evaluation. {question} is the original instruction; {answer 1} is the original response, and {answer 2} is the response generated by the LVLM using MARINE.

Prompt template for GPT-4V-aided evaluation

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image.

Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:

1. Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.
2. Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.

Please output a single line for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Question: {question}

Assistant 1: {answer 1}

Assistant 2: {answer 2}

Output format:

Accuracy:

Scores of the two answers:

Reason:

Detailedness:

Scores of the two answers:

Reason:

a maximum token length of 64, without any stopping criteria, using a single A6000 GPU. Then latency was calculated as the ratio of the number of output tokens and encoding and generation time.

Table 16. Detailed POPE (Li et al., 2023b) results on three datasets (MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), GQA (Hudson & Manning, 2019)).

Dataset	Type	Model	w/MARINE	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Yes(%)
MSCOCO	Adversarial	LLaVA	✗	51.8	50.9	99.5	67.4	97.7
			✓	66.9	61.7	89.1	72.9	72.3
		mPLUG-Owl2	✗	72.5	65.5	94.9	77.5	72.4
	Popular		✓	82.8	83.4	82.0	82.7	49.2
		LLaVA	✗	52.4	51.2	99.8	67.7	97.4
			✓	71.3	65.8	88.9	75.6	67.5
	Random	mPLUG-Owl2	✗	75.8	68.7	94.9	79.7	69.0
			✓	85.6	88.4	82.0	85.1	46.4
		LLaVA	✗	58.3	54.5	99.7	70.5	91.4
	Random		✓	78.5	73.4	89.3	80.6	60.8
		mPLUG-Owl2	✗	81.8	75.2	94.9	83.9	63.1
			✓	88.1	93.4	81.9	87.3	43.9
A-OKVQA	Adversarial	LLaVA	✗	50.0	50.0	99.5	66.6	99.5
			✓	56.3	53.6	94.3	68.3	88.1
		mPLUG-Owl2	✗	62.5	57.3	98.1	72.3	85.6
	Popular		✓	74.4	68.8	89.3	77.7	64.9
		LLaVA	✗	50.1	50.1	99.8	66.7	99.7
			✓	63.0	58.0	94.5	71.9	81.6
	Random	mPLUG-Owl2	✗	69.1	62.1	97.9	76.0	78.9
			✓	82.5	78.8	89.1	83.6	56.5
		LLaVA	✗	55.4	52.8	99.8	69.1	94.4
	Random		✓	73.7	66.7	94.7	78.3	71.0
		mPLUG-Owl2	✗	77.2	69.2	98.2	81.2	71.0
			✓	89.2	89.2	89.3	89.2	50.1
GQA	Adversarial	LLaVA	✗	50.3	50.1	99.8	66.8	99.5
			✓	54.4	52.5	93.8	67.3	89.4
		mPLUG-Owl2	✗	68.4	63.0	98.2	75.6	79.8
	Popular		✓	76.0	73.6	81.2	77.2	55.2
		LLaVA	✗	50.1	50.0	99.8	66.7	99.7
			✓	58.7	55.1	94.3	69.5	85.5
	Random	mPLUG-Owl2	✗	70.6	63.8	94.9	76.3	74.4
			✓	77.6	75.6	81.3	78.4	53.8
		LLaVA	✗	55.7	53.0	99.8	69.2	94.1
	Random		✓	74.3	67.3	94.8	78.7	70.5
		mPLUG-Owl2	✗	82.0	75.2	95.5	84.1	63.5
			✓	86.8	91.5	81.3	86.1	44.4

Table 17. Performance on general metrics for the image captioning task, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) scores(%).

Model	w/MARINE	BLEU_1 (↑)	BLEU_2 (↑)	BLEU_3 (↑)	BLEU_4 (↑)	ROUGE_L (↑)	CIDEr (↑)	SPICE (↑)
LLaVA	✗	14.06	7.12	3.72	1.90	22.06	0.08	16.77
	✓	18.59	9.96	5.47	3.04	26.02	0.21	20.58
mPLUG-Owl2	✗	39.91	25.16	16.57	11.24	36.26	1.05	26.82
	✓	39.51	24.37	15.93	10.70	36.01	1.03	27.42

B Additional Experiments

B.1 Additional Baselines

To further contextualize the effectiveness of MARINE, we conducted additional experiments comparing our approach to a baseline that employs carefully engineered prompts designed to reduce hallucination. Specifically, we used the following prompt:

Table 18. Performance on general metrics for the LLaVA-QA90 task, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) scores(%).

Model	w/MARINE	BLEU_1 (\uparrow)	BLEU_2 (\uparrow)	BLEU_3 (\uparrow)	BLEU_4 (\uparrow)	ROUGE_L (\uparrow)	CIDEr (\uparrow)	SPICE (\uparrow)
LLaVA	\times	21.02	12.91	8.79	6.41	32.30	0.93	31.36
	\checkmark	23.37	14.39	9.59	6.83	33.81	0.99	31.91
mPLUG-Owl2	\times	44.50	28.57	19.58	14.43	40.24	1.46	40.51
	\checkmark	45.82	28.87	19.24	13.70	38.54	1.29	38.70

Table 19. Comparison against carefully engineered prompts.

Method	LLaVA			LLaVA-v1.5			mPLUG-Owl2		
	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow
CHAIR									
Original	26.6	10.5	47.4	8.8	4.6	41.1	5.0	3.2	33.2
Direct Prompting	27.2	11.0	46.4	19.6	8.3	52.3	9.0	5.1	42.0
Prompts as Additional Guidance	37.4	10.5	50.4	12.6	5.9	44.6	6.6	3.9	40.4
MARINE (ours)	17.8	7.2	50.8	6.2	3.0	44.3	4.2	2.3	41.4

Table 20. Experiments on dynamic guidance strength based on confidence scores on CHAIR metrics.

Method	LLaVA			mPLUG-Owl2		
	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow
CHAIR						
Fix Guidance Strength	17.8	7.2	50.8	4.2	2.3	41.4
Dynamic Guidance Strength	14.8	6.5	49.9	5.0	2.6	41.0

Table 21. Experiments on dynamic guidance strength based on confidence scores on POPE metrics.

Method	LLaVA			mPLUG-Owl2		
	Accuracy \uparrow	F1 \uparrow	Yes Ratio	Accuracy \uparrow	F1 \uparrow	Yes Ratio
POPE						
Fix Guidance Strength	66.9	72.9	72.3	82.8	82.7	49.2
Dynamic Guidance Strength	71.97	74.48	59.83	83.3	83.2	49.4

Describe the visible contents of this image in as much detail as possible without adding any information not clearly visible. Only mention objects, colors, shapes, and textures that can be directly observed in the image, avoiding assumptions about materials, functions, or contexts. If there are any uncertainties about what an object is, describe its visual characteristics (e.g., ‘a circular object with a smooth surface’) without inferring its purpose or identity. Avoid creative or hypothetical descriptions, and focus on observable details only.

With two different settings:

- *Direct Prompting*: The original input query was replaced with the prompts as described.
- *Prompts as Additional Guidance*: We incorporated the prompt as supplemental context to guide the models in generating outputs.

As shown in Table 19, prompt-based guidance can improve recall for some models (e.g., LLaVA-v1.5), but does not consistently reduce hallucinations across all metrics. In fact, CHAIR scores often worsen. In contrast, MARINE achieves stronger improvements across all models.

We highlight two key differences between MARINE and prompt-based approaches:

- *Model Dependence*: Prompting methods rely heavily on the instruction-following capabilities of the model. While they may reduce hallucinations slightly for stronger models (e.g., LLaVA-v1.5), they can worsen performance in weaker models (e.g., LLaVA). Additionally, prompt-based approaches may require fine-tuning to be effective (Deng et al., 2024). MARINE, by contrast, improves grounding through explicit visual signals, making it effective even without model retraining.
- *Generalization and Efficiency*: Prompting methods often require task-specific tuning or dataset-aware phrasing. MARINE generalizes across tasks and models with minimal engineering and no fine-tuning, while offering more consistent hallucination reduction.

B.2 Dynamic Guidance Strength

We conducted additional experiments to compare fixed and dynamic guidance strength strategies using both CHAIR and POPE metrics (Tables 20 and 21).

- *Fix Guidance Strength* uses a fixed guidance strength of 0.7, selected to balance hallucination reduction and instructions adherence.
- *Dynamic Guidance Strength* adjusts the guidance strength dynamically by mapping the mean confidence score (s) of the image-grounding models to a range of (0.4, 0.8) using the formula

$$\gamma' = 0.4 + \frac{(0.8 - 0.4) \cdot (s - s_{\min})}{s_{\max} - s_{\min}}.$$

A higher confidence score indicates more reliable grounding, which results in stronger guidance. Empirically, we find that dynamic guidance improves performance for weaker models such as LLaVA, which are more sensitive to noisy signals. For stronger models like mPLUG-Owl2, a fixed guidance strength is already sufficient to reduce object hallucinations effectively.

B.3 Effect of Sampling Temperature

In our main experiments, we use greedy decoding (temperature = 0) to ensure deterministic outputs and reproducible comparisons—consistent with our primary baseline (VCD) and common practice in hallucination benchmarks. To evaluate robustness under stochastic decoding, we also test with a temperature of 0.6 and report mean \pm standard deviation in Table 22. MARINE continues to outperform baseline generations across all hallucination metrics, demonstrating effectiveness regardless of sampling strategy.

Table 22. Object hallucination metrics under temperature = 0.6 sampling.

Method	LLaVA			mPLUG-Owl2		
	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow	$C_s \downarrow$	$C_i \downarrow$	Recall \uparrow
CHAIR						
Greedy	26.1 ± 1.6	10.8 ± 0.5	46.0 ± 0.8	4.9 ± 0.6	2.8 ± 0.3	37.7 ± 0.6
MARINE (ours)	19.3 ± 0.8	7.6 ± 0.1	50.6 ± 0.2	4.5 ± 0.6	2.4 ± 0.2	41.1 ± 0.4

B.4 Memory Analysis

We evaluated the peak GPU memory usage during inference on 500 image captioning examples using the LLaVA model, with a batch size of 16 and a maximum generation length of 64 tokens. Results are reported in Table 23. Although MARINE introduces additional vision models, the overall memory footprint increases by only approximately 30% during inference—significantly less than doubling. This is because the object detection models are relatively lightweight compared to the large LLM backbone.

B.5 Further Study on Guidance Strength

Figure 5 shows how varying the guidance strength γ affects the quality of LLaVA’s output on the LLaVA-QA90 and image captioning tasks (max generation length = 256). We observe that setting $\gamma = 1$ does not yield the best image captioning performance. In the LLaVA-QA90 task, guidance strengths in the range of 0.5 to 0.7 lead to higher output quality. This observation is consistent with prior findings in classifier-free guidance literature: overly strong guidance can dominate the generation process and reduce fluency or instruction adherence.

To further validate these results, we use GPT-4V as an automatic judge to score outputs (on a 10-point scale) for accuracy and detail. The results, summarized in Table 24, show that balancing the original LVLM branch leads to improved generation quality. Finally, Figure 6 provides qualitative examples showing how excessive guidance can reduce instruction alignment, often introducing unnecessary visual details into the response.

Table 23. Peak GPU Memory Usage during Inference (GB) of MARINE compared to greedy decoding and VCD.

Metric	Greedy	VCD	MARINE (Ours)
Peak GPU Memory Usage	23.53	20.73 ($\times 0.88$)	30.78 ($\times 1.30$)

Table 24. Results of GPT-4V-aided evaluation. The accuracy and detailedness metrics are on a scale of 10, and a higher score indicates better performance. The symbols \times and \checkmark indicate performance metrics without and with our method, respectively.

Task	Metric \uparrow	\times ($\gamma = 1$)	\checkmark ($\gamma = 0.7$)
LLaVA-QA90	Accuracy	5.52	5.79
	Detailedness	4.58	4.77
Image Captioning	Accuracy	6.06	6.22
	Detailedness	5.00	5.24

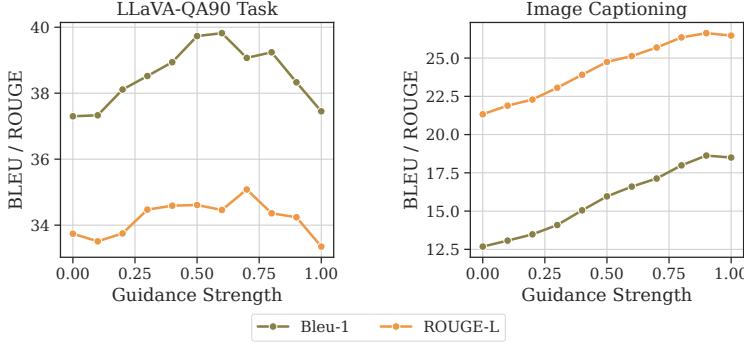


Figure 5. The impact of guidance strength on the output text quality.

C Further Analysis

C.1 Limitations of Hallucination Evaluation

While CHAIR and POPE are widely adopted for evaluating object hallucinations in vision-language models, both have inherent limitations. CHAIR depends on a fixed object vocabulary and synonym list, which may miss rare or fine-grained concepts. POPE relies on the quality of segmentation tools to define ground-truth objects, introducing variability across settings.

To address these limitations, we incorporate *ALOHa* (Automatic Localized Hallucination) (Petryk et al., 2024), a reference-based metric that evaluates hallucination at both the object level (*ALOHa*₀) and the caption level (*ALOHa*). We follow the standard ALOHa setup using MSCOCO ground-truth captions and enable reference object detection for more precise and generalizable assessment. As shown in Table 25, MARINE consistently outperforms greedy decoding across all models and both ALOHa metrics.

C.2 Additional Related Work

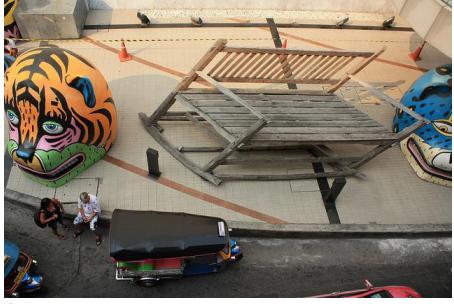
Several recent works aim to improve grounding or reduce hallucination in vision-language models. BRAVE (Kar et al., 2024) enhances faithfulness by combining diverse visual sources, similar in spirit to MARINE, but introduces additional trainable modules. MARINE achieves comparable performance with a training-free, modular design.

Other approaches focus on evaluation (Hu et al., 2023; Cho et al., 2024; Lin et al., 2024) or feature-level interventions (Yang et al., 2025; Liu et al., 2024a) to steer models away from hallucinations. Liu et al. (2024b) address *text inertia*, where models generate similar outputs regardless of image content. Wan et al. (2024) introduce sub-image contrastive alignment, and Zhang et al. (2024b) control generation by adjusting visual attention weights.

These methods highlight complementary strategies to MARINE’s structured, object-level guidance for reducing hallucination.

C.3 Effect of MARINE on logit distribution.

In Figure 7, we illustrate a specific example that shows how MARINE influences the logit distribution of LVLMs during text generation. Specifically, MARINE is observed to selectively target the potential hallucinated tokens, reducing their original probabilities to mitigate the risk of hallucination in the generated text. For instance, in the provided example, the probability of “fork” is significantly lowered with MARINE, which would have originally resulted in a hallucinated object.



GPT4 response as reference

The art installation, featuring an oversized wooden bench and two massive animal art sculptures, encourages interaction among people by providing a visually intriguing and inviting space to sit or gather. People are naturally drawn to the unique and whimsical nature of these installations, sparking curiosity and conversation among those who encounter them. The oversized bench provides seating for multiple individuals, prompting people to sit down, rest, or socialize while admiring the giant painted feline heads and other artworks on display. Additionally, the installations capture the attention of passersby, creating a shared experience and a sense of community among those who visit the area. The artistic environment stimulates discussion and appreciation of the art, facilitating social interactions and connections among those who gather near the sculptures and oversized bench.

Question: How might the art installation encourage interaction among people?



LLaVA w/ MARINE with balance guidance

The art installation, which consists of a large, colorful, and intricately painted tiger sculpture, is likely to encourage interaction among people in several ways. Firstly, the vibrant colors and unique design of the tiger sculpture can attract attention and draw people towards it, making them curious about its origin and significance. This curiosity can lead to conversations and discussions among passersby, fostering social interactions and a sense of community.\n\nSecondly, the sculpture's size and prominence in the city's public space can encourage people to take a break from their daily routines and engage with the artwork. The presence of the tiger sculpture can create a sense of awe and wonder, encouraging people to pause and appreciate the beauty of the artwork. This can lead to a more relaxed and enjoyable atmosphere, as people take the time to admire the art and enjoy the surrounding environment.\n\nLastly, the sculpture's placement in the city's public space can also serve as a focal point for community events, such as art exhibitions, cultural festivals, or even street performances. The tiger sculpture can become a gathering [reach the maximum generation length]

LLaVA w/ MARINE with overly strong guidance

The art installation, which features a large, colorful tiger sculpture and a car parked next to it, can encourage interaction among people in several ways. Firstly, the vibrant colors and unique design of the tiger sculpture can attract attention and draw people towards it, making them curious about its origin and purpose. This curiosity can lead to conversations and discussions among passersby, encouraging social interaction and engagement. Additionally, the presence of the car parked next to the sculpture can provide a focal point for people to gather and take photos, further fostering social interaction and creating a sense of community. The art installation can also serve as a backdrop for events or gatherings, such as art exhibitions, festivals, or even photo shoots, which can further encourage people to interact with each other and engage with the artwork. Overall, the art installation can serve as a catalyst for social interaction and community engagement, promoting a sense of connection and shared experience among people.

Figure 6. This case highlights that overly strong guidance can induce the model to prioritize providing exhaustive visual details from the image, even when such details are irrelevant to the specific instruction (e.g., “a car parked next to it”). In contrast, balanced guidance enables the model to maintain better adherence to the instruction while still utilizing the visual information effectively.

Conversely, standard language elements such as “various”, an adjective describing the overall image context, and “with”, a crucial preposition, maintain their original probabilities. This selective nature of modulation by MARINE ensures coherent and contextually relevant text generation that adheres to the instruction while effectively reducing hallucinations.

C.4 Discussion on fine-tuning methods.

The examples shown in Figure 8 illustrate that LURE, at times, fails to adhere to the given instructions when correcting LVLM generations. Despite receiving concise image descriptions generated based on instructions for short responses, LURE predominantly overwrites them with excessively long responses that contain information irrelevant to the instruction. Furthermore, LURE fails to adequately address the binary question format of POPE, as LURE fixates on extended descriptions without responding with “yes” or “no”, making its evaluation using POPE impractical. This issue can be prevalent in small-scale fine-tuning methods, where the limited variety of the specifically tailored fine-tuning dataset harms the model’s performance on other tasks. In contrast, the training-free approach of MARINE demonstrates effective mitigation of hallucinations across a variety of question formats.

C.5 Extended Analysis in Ablation Study

Additional experimental results explore the score threshold of object grounding features, which are examined across LLaVA, and mPLUG-Owl2, with findings presented in Figures 9, and 10.

This variation is achieved by implementing four confidence thresholds (0.5, 0.7, 0.9, and 0.95) in the DETR model predictions (with MARINE-Truth serving as an ideal reference), where higher thresholds correspond to lesser, yet higher-

Table 25. ALOHa hallucination scores (all values are in %). MARINE improves over greedy decoding across models and metrics.

Method	LLaVA		LLaVA-v1.5		mPLUG-Owl2	
	ALOHa ↑	ALOHa ₀ ↑	ALOHa ↑	ALOHa ₀ ↑	ALOHa ↑	ALOHa ₀ ↑
Greedy	40.1	70.1	61.9	83.1	70.2	87.0
MARINE	48.7	76.1	66.7	85.6	72.9	88.2

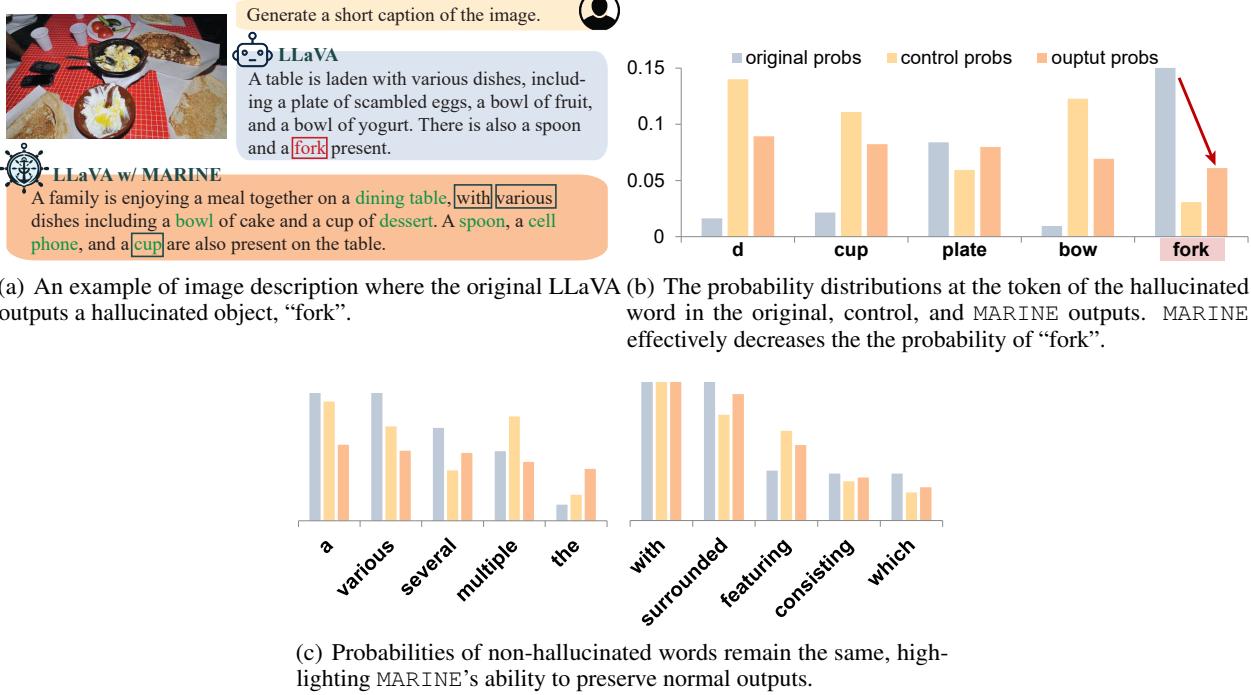


Figure 7. This sample shows how MARINE controls logit distributions to mitigate hallucinations like “fork” while preserving the probabilities of “with”, “various” during generation.

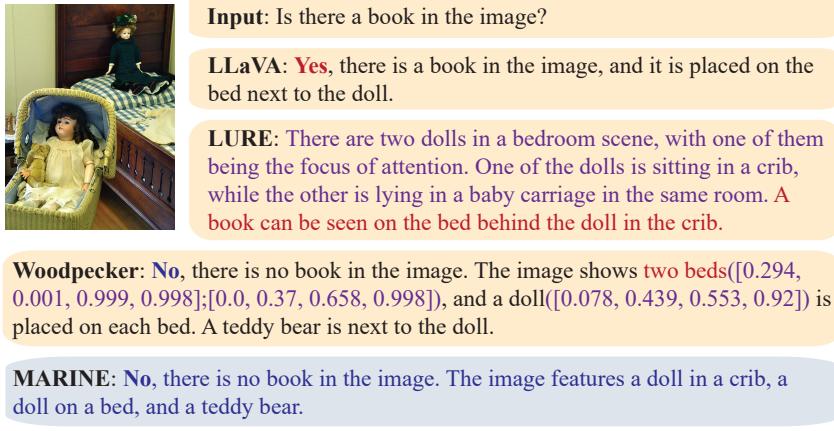


Figure 8. Example responses to an image-question pair. The LURE-corrected output deviates from the original question, offering irrelevant descriptions without directly addressing the query. Woodpecker hallucinates the existence of two beds while there is only one bed in the figure. In contrast, MARINE maintains the original answer’s style and adheres to the user’s instruction while eliminating hallucination.

quality, visual information. Our findings highlight two significant insights. Firstly, an increase in the quality of visual information correlates with a noticeable decrease in hallucinations produced by the LVLMs. A lower threshold, which allows for more visual information but also includes noisier content, could potentially result in an increased occurrence of hallucinations. Furthermore, lower-quality visual information is associated with enhanced Recall. This suggests that LVLMs under guidance, despite the presence of noisy visual inputs, tend to focus more on the visual details (i.e., objects),

resulting in more elaborate descriptions.

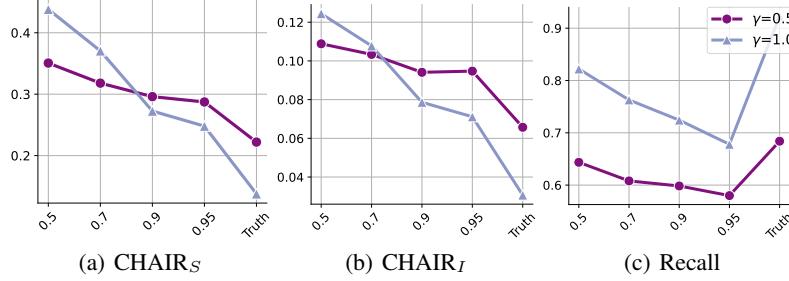


Figure 9. LLaVA’s performance on CHAIR according to different score threshold of object grounding features in MARINE. We consider four confidence thresholds (0.5, 0.7, 0.9, and 0.95) for DETR to vary the score threshold.

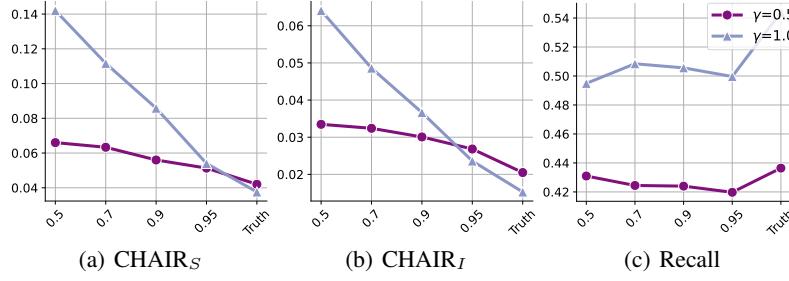


Figure 10. mPLUG-Owl2’s performance on CHAIR according to different score threshold of object grounding features in MARINE. We consider four confidence thresholds (0.5, 0.7, 0.9, and 0.95) for DETR to vary the score threshold, with MARINE-Truth serving as an ideal reference.

C.6 More Case Studies

In Figures 11, 12 and 13, we present examples of the outputs from LURE (Zhou et al., 2023), Woodpecker (Yin et al., 2023) and MARINE on different tasks further validate our arguments in the paper.



Query: What is the position of the skateboard in the image?

LLaVA: The skateboard is on the ground, with the rider standing on it and holding onto the trucks.

LLaVA w/ MARINE: The skateboard is on the ground, with the person standing on it.

Query: Generate a short caption of the image.

LLaVA-v1.5: A family of four, including a woman and three children, are sitting at a table with a plate of food in front of them.



LLaVA-v1.5 w/ MARINE: A family of four, including two adults and two children, are sitting around a dining table with plates of food in front of them.

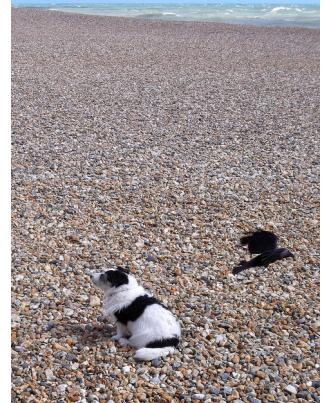
Figure 11. Hallucination mitigation examples by our proposed MARINE across multiple tasks: LLaVA-QA90 and image captioning. Hallucinated objects generated by the LVLM are highlighted in red.

Query: Generate a short caption of the image.

LLaVA: A black and white dog is sitting on a beach, surrounded by rocks and sand. The dog is looking at its owner who is standing behind it, as if waiting for a command or attention. The scene captures the dog's loyalty and affection towards its owner.

(Note: 'bird' omitted)

LURE: The scene features a black and white dog sitting on a beach, surrounded by pebbles. The dog is looking at its owner who is standing behind it, as if waiting for attention. The background reveals a vast ocean, and other birds can be seen scattered throughout the area. The presence of these birds gives the impression that the dog is not alone, and they are enjoying the time together on the beach. Overall, the image captures a charming scene on the sandy beach.



Woodpecker: A black and white dog([0.243, 0.662, 0.551, 0.894]) is sitting on the beach([0.002, 0.029, 0.998, 0.999]), surrounded by sand. The dog is looking at its owner([0.712, 0.57, 0.92, 0.684]) who is standing behind it.

LLaVA w/ MARINE:

In the image, a black and white dog is standing on a beach, looking at a bird.

Figure 12. A comparison of responses from baseline models and our MARINE in an image description task. It illustrates MARINE's superior ability to reduce hallucinations, in contrast to LURE and Woodpecker, which fail to effectively address hallucinations and sometimes even increase hallucinated content. This example highlights the strengths of our correct-during-generation framework over post-correction approaches, showcasing its efficiency, preservation of original style, and enhanced adherence to instructions.

Query: Generate a short caption of the image.

MiniGPT-v2: no

Woodpecker: No, there is no toothbrush in the image.

MiniGPT-v2 w/ MARINE: yes



Figure 13. A comparison of responses from baseline models and our MARINE in POPE “yes-or-no” task. MiniGPT-v2 provides a concise response without referencing any objects. Under these circumstances, Woodpecker is unable to perform corrections via GPT-3.5 due to missing visual details. MARINE, however, successfully corrects the response while retaining MiniGPT-v2’s style.