SAVER: Mitigating Hallucinations in Large Vision-Language Models via Style-Aware Visual Early Revision

Zhaoxu Li[1,2], Chenqi Kong[2*], Yi Yu[2], Qiangqiang Wu[3], Xinghao Jiang[4], Ngai-Man Cheung[5], Bihan Wen[2], Alex Kot[2,6], Xudong Jiang[2]

[1]ROSE Lab, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore
[2]ROSE Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3]City University of Hong Kong, Hong Kong SAR
[4]Shanghai Jiao Tong University, China
[5]Singapore University of Technology and Design, Singapore
[6]VinUniversity, Hanoi, Vietnam
{zhaoxu001, yuyi0010}@e.ntu.edu.sg, qiangqwu2-c@my.cityu.edu.hk, xhjiang@sjtu.edu.cn, ngaiman_cheung@sutd.edu.sg, {chenqi.kong, bihan.wen, eackot, exdjiang}@ntu.edu.sg

Abstract

Large Vision-Language Models (LVLMs) recently achieve significant breakthroughs in understanding complex visual-textual contexts. However, hallucination issues still limit their real-world applicability. Although previous mitigation methods effectively reduce hallucinations in photographic images, they largely overlook the potential risks posed by stylized images, which play crucial roles in critical scenarios such as game scene understanding, art education, and medical analysis. In this work, we first construct a dataset comprising photographic images and their corresponding stylized versions with carefully annotated caption labels. We then conduct head-to-head comparisons on both discriminative and generative tasks by benchmarking 13 advanced LVLMs on the collected datasets. Our findings reveal that stylized images tend to induce significantly more hallucinations than their photographic counterparts. To address this issue, we propose **S**tyle-**A**ware **V**isual **E**arly **R**evision (**SAVER**), a novel mechanism that dynamically adjusts LVLMs' final outputs based on the token-level visual attention patterns, leveraging early-layer feedback to mitigate hallucinations caused by stylized images. Extensive experiments demonstrate that SAVER achieves state-of-the-art performance in hallucination mitigation across various models, datasets, and tasks.

Extended version —
https://www.arxiv.org/abs/2508.03177

Introduction

Large Vision-Language Models (LVLMs) (OpenAI 2023; Chen et al. 2023; Yin et al. 2024a; Zhao et al. 2024) have achieved remarkable successes in a plethora of applications in the past few years (Li et al. 2022; Liu et al. 2023a; Zhu et al. 2023). However, the phenomenon of hallucination can cause severe consequences in some practical scenarios, such as medical analyzes (Hu et al. 2023; Wang et al. 2023b), autonomous driving (Chen et al. 2024a; Liu et al. 2023b), and human-computer interaction (Brie et al. 2023). These issues raise pressing security concerns and may cause unintended harm to the public, significantly hindering the real-world deployment of LVLMs.

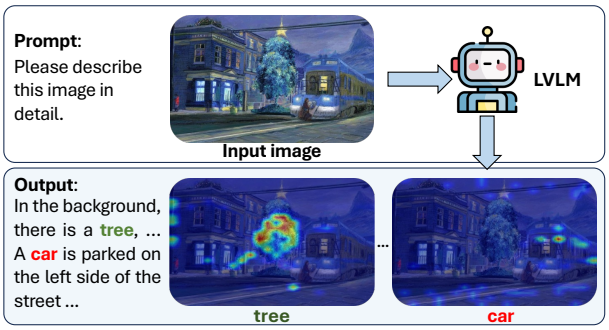

Figure 1: Correlation map between image tokens and generated tokens. Red: hallucinated tokens, showing sparse, low-confidence correlations. Green: real tokens, showing concentrated clusters over the corresponding regions.

To mitigate hallucinations, existing approaches can be broadly categorized into two groups: instruction tuning methods (Sun et al. 2023; Xing et al. 2024) and decoding-based methods (Chuang et al. 2023; Jiang et al. 2024). The former requires retraining LVLMs using curated datasets and tuning strategies, which effectively reduce hallucinations while introducing expensive computational costs. The latter employs post-hoc correction mechanisms, adjusting logit scores at each generative step to encourage LVLMs to focus more on the corresponding visual content. While these methods have shown promising results for photographic images, they suffer from significant performance drops when applied to stylized images (e.g., game and sketch).

Stylized images play a critical role in various applications, including art education, medical analysis, and criminal forensics. For instance, LVLMs can interpret forensic sketches to retrieve matches from photographic databases

*Corresponding author

or generate suspect profiles. Minimizing hallucinations in such scenarios is essential to enhance the reliability of these models. However, a comprehensive benchmark for assessing LVLM performance on stylized images remains unavailable. To fill this gap, we construct a dataset comprising paired captions and both photographic and style-transferred images generated using state-of-the-art methods across five stylistic domains. We evaluate 13 advanced LVLMs and observe that stylized images significantly increase hallucination rates compared to their original photographic counterparts.

To investigate the underlying causes, we conduct an empirical analysis by measuring the correlations between stylized image representations and the generated tokens. Fig. 1 visualizes the representational patterns of real and hallucinated objects in stylized images. We observe that real object representations exhibit concentrated, high-confidence activation regions, while the hallucinated object displays sparse, low-confidence distributions. This insight naturally inspires the design of a decoding strategy that encourages the model to focus more precisely on the relevant visual regions.

We propose **SAVER**: **S**tyle-**A**ware **V**isual **E**arly **R**evision, a training-free hallucination mitigation strategy designed to address hallucination issues in stylized images. Due to the strong knowledge priors of language models, LVLMs tend to progressively suppress visual information in later model layers (Wang et al. 2024a; Jiang et al. 2024). To counteract this, SAVER dynamically searches the optimal preceding layers with highly concentrated representational patterns to refine the output token logits. As a plug-and-play decoding strategy, SAVER can be seamlessly integrated into various LVLMs and significantly reduces hallucination rates for stylized images without requiring additional training.

Our key contributions are: (1) To the best of our knowledge, we are the first to construct a captioning dataset specifically for stylized images. We further establish a benchmark using 13 advanced LVLMs and find that stylized images tend to produce more hallucinations; (2) Compared to hallucinated tokens, we demonstrate that the correlation patterns between correct tokens and input visual contents exhibit denser visual activations in early layers. Based on this insight, we propose SAVER, a training-free decoding strategy that dynamically corrects generated tokens by leveraging high-confidence activations from earlier layers; (3) Extensive experimental results show that SAVER consistently outperforms previous methods, effectively mitigating hallucinations across a variety of models, datasets, and tasks.

Related Work

Large Vision-Language Models (LVLMs). Large Language Models (LLMs) such as LLaMA (Touvron et al. 2023) and Vicuna (Chiang and Li 2 May 2025) have achieved remarkable advancements. The rapid development of LVLMs has significantly enhanced the ability of foundation models to interpret and reason about visual content. Early LVLMs, including BLIP (Li et al. 2022) and LLaVA (Liu et al. 2023a), extended LLMs to handle image understanding and reasoning tasks. To bridge the modality gap between vision and language, various approaches have been adopted, such as linear projection layers (e.g., LLaVA (Liu et al. 2023a), MiniGPT-4 (Zhu et al. 2023)), Q-former modules (e.g., BLIP-2 (Li et al. 2023a), InstructBLIP (Dai et al. 2023)), and cross-attention mechanisms (e.g., Flamingo (Alayrac et al. 2022), OpenFlamingo (Awadalla et al. 2023)). More recent models, such as GPT-4o (OpenAI 2023) and Gemini (AI 2024), have demonstrated exceptional capabilities in visual reasoning and understanding. Nevertheless, despite these advancements, hallucination remains a significant challenge.

Visual Hallucination in LVLMs typically refers to cases where the generated text is inconsistent with the input image at the instance level (Rohrbach et al. 2018; Zhou et al. 2023; Li et al. 2023b) or the faithfulness of the generated free-form answer (Jing et al. 2023). There are various potential factors that can cause hallucinations, including modality gap, training data bias, error accumulation, etc. (Zhou et al. 2023). To evaluate LVLM hallucination levels, CHAIR (Rohrbach et al. 2018) proposes measuring object hallucination rates in output captions. The POPE benchmark (Li et al. 2023b) assesses object hallucinations using binary "Yes/No" questions. Additionally, MME (Fu et al. 2024) provides a more challenging dataset for hallucination evaluation, which encompasses various hallucination types, such as object, attribute, counting, etc. AMBER (Wang et al. 2023a) supports both generative and discriminative tasks, covering existence, attribute, and relation hallucinations. While prior works have primarily focused on natural photographic images, this study constructs a new dataset and comprehensive benchmark to investigate hallucination in LVLMs when processing stylized images.

Hallucination Mitigation. Existing methods fall into tuning-based and decoding-based categories. Tuning-based approaches curate specialized datasets (Wang et al. 2024c; Zhang et al. 2024; Jing and Du 2024) and apply alignment training (Xie et al. 2024; Sun et al. 2023; Xing et al. 2024), but are costly in annotation and computation. In contrast, training-free methods (Yin et al. 2024b) and decoding-based techniques (Huang et al. 2024; An et al. 2024) are more efficient. Recent contrastive decoding methods (Leng et al. 2024; Chen et al. 2024c; Chuang et al. 2023) further enhance performance by leveraging visual comparisons. Deco (Wang et al. 2024a) and Attention Lens (Jiang et al. 2024) emphasize layer selection to mitigate hallucination. Our method, **SAVER**, differs from these two approaches by dynamically selecting non-hallucinated tokens without the need to train a detector. And SAVER leverages richer visual signals during layer selection for final output revision, thereby achieving better hallucination mitigation performance.

Benchmarking Stylized Image Hallucination

In this section, we introduce our proposed style dataset by outlining its motivation, construction process, and image generation pipeline. We then present the benchmark along with a detailed description of the procedures. An example style image from the dataset is shown in Fig. 2.

Motivation. We propose this dataset to examine object hallucination in cross-style scenarios, evaluating LVLMs' ability to analyze images from diverse domains. This section describes the construction of our style-diverse dataset and

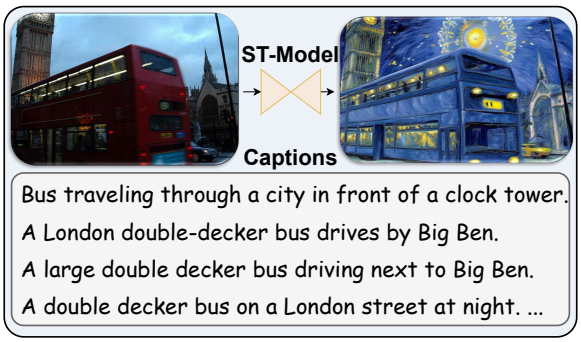

Figure 2: Top left: original image; top right: stylized image generated by Style Transfer (ST) model; bottom: COCO captions listing all salient objects.

benchmark. Existing LVLM benchmarks mainly use photographic ("Original") images, reflecting their dominance in real-world applications. However, images also come in styles like Cartoon and Sketch, where objects may appear visually distinct (e.g., a cat with unusual colors). As shown in Fig. 2, such variations pose challenges for LVLMs, which may lack exposure to these styles during training. To address this, we introduce a framework to assess hallucination behaviors across a range of artistic styles.

Dataset. Follow previous hallucination works (Rohrbach et al. 2018; Li et al. 2023b), we sampled images from the COCO dataset to construct a high-quality stylized dataset. Each selected image contains at least five annotated objects to create a challenging evaluation setting. We applied SOTA InstantStyle (Wang et al. 2024b) model to generate style-transferred images in five styles: Cartoon, Game, Graffiti, Painting, and Sketch, ending up with 1,800 images. We carefully checked each pair of images and manually filter out the low-quality one, ensuring that each original image and its stylized versions share identical annotations, including object and caption labels.

Benchmark. We construct a benchmark using two metrics:

- **CHAIR:** Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al. 2018) measures the proportion of hallucinated objects—those mentioned in the caption but absent in the image. We use two variants: CHAIRi (instance-level) and CHAIRs (sentence-level), defined as:

$$\text{CHAIR}_i = \frac{\{\text{hallucinated instances}\}}{\{\text{all mentioned instances}\}},$$
$$\text{CHAIR}_s = \frac{\{\text{captions with hallucinations}\}}{\{\text{all captions}\}}. \quad (1)$$

- **POPE:** Polling-based Object Probing Evaluation (POPE) (Li et al. 2023b) evaluates hallucination by asking LVLMs "Yes/No" questions about object presence, using three negative sampling strategies: random, popular (top-k frequent absent objects), and adversarial (top-k absent objects ranked by co-occurrence with ground-truth). We follow the original POPE settings and generate 6 questions for each image, resulting in 1800 $\times$ 6 $\times$ 3 question image pairs.

Style Aware Visual Early Revision

We conduct a comprehensive benchmark and study to uncover the underlying mechanisms that lead to object hallucination when processing stylized images. Guided by these findings, we propose **S**tyle-**A**ware **V**isual **E**arly **R**evision (**SAVER**), a novel inference-time strategy aimed at reducing hallucinations caused by style-transferred content. The overall design and workflow of SAVER are illustrated in Fig. 3.

Preliminaries

LVLMs for text generation typically consist of three key components: a vision encoder, a projection module, and an autoregressive language model. The vision encoder first converts an input image into a sequence of visual tokens $X_V = \{x_{v_1}, x_{v_2}, \ldots, x_{v_P}\}$. In parallel, a text prompt is tokenized into Q textual tokens $X_C = \{x_{c_1}, x_{c_2}, \ldots, x_{c_Q}\}$. Here P and Q are the lengths of the visual and textual tokens. The visual and textual embeddings are concatenated to form the model input X, which is passed through an autoregressive language model composed of N stacked transformer decoder layers. At each layer i, the model produces hidden states $h^i = \{h_0^i, h_1^i, \ldots, h_{T-1}^i\}$, where $T = P + Q$, P is the length of the visual tokens. During generation, the hidden state at the final position h_{T-1}^N is projected via an affine transformation $\phi(\cdot)$, typically using an unembedding matrix $W_U \in \mathbb{R}^{|V| \times d}$, to produce a logit distribution over the vocabulary V.

Stylization Amplifies Object Hallucination

We investigate the impact of image stylization on object hallucination in LVLMs, specifically in the context of image captioning. Our benchmark includes various LVLMs, covering both open-source models (e.g., LLaVA (Liu et al. 2023a), MiniGPT-4 (Zhu et al. 2023), InstructBLIP (Dai et al. 2023), TinyLLaVA (Zhou et al. 2024), Phi-3-V (Abdin et al. 2024), Fuyu (Bavishi et al. 2023), Idefics2 (Laurençon et al. 2024), mPLUG-Owl2 (Ye et al. 2024),Qwen-VL (Bai et al. 2023)) and closed-source platforms (e.g., GPT-4o (OpenAI 2023), Gemini-1.5-Pro (AI 2024)).

To elicit diverse responses, we employ two prompt templates: a detailed prompt (*"Please describe this image in detail"*) and a concise prompt (*"Provide a one-sentence caption for the provided image."*). The generated captions are evaluated using the CHAIR metric, with results summarized in Tab. 1 and Tab. 7 (Appendix). In addition, we assess hallucination across three splits of the Style-POPE benchmark, with detailed results visualized in Fig. 5 in the Appendix.

Our experiments reveal a consistent increase in object hallucination when models process stylized images, such as those rendered in Cartoon, Game, Graffiti, Painting, and Sketch styles, compared to original photographs. This performance degradation highlights the models' reduced grounding capabilities under stylistic shifts.

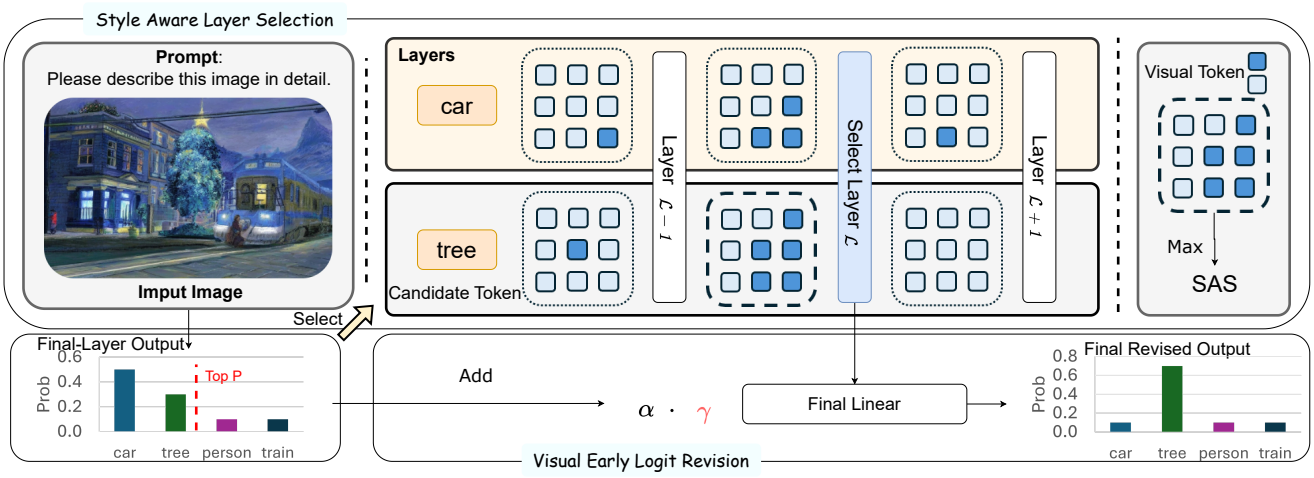

Figure 3: SAVER first selects the top-p tokens, then chooses a layer via the Style-Aware Score, and finally revises the final-layer output. Darker visual tokens indicate higher confidence.

Model	Cartoon		Game		Graffiti		Painting		Sketch		Original		Average	
	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs
GPT4o	10.3	29.0	7.5	23.0	7.9	21.3	6.7	20.0	5.6	17.7	4.7	16.7	7.1	21.3
Gemini-1.5-PRO	13.5	25.3	7.9	14.3	13.1	18.7	11.3	19.3	7.0	13.3	6.2	18.3	9.8	18.2
LLaVA-1.5	12.0	43.3	9.1	31.7	11.4	37.7	11.2	34.7	11.4	37.7	6.8	26.7	10.3	35.3
LLaVA-1.5-13b	10.7	38.7	10.3	35.0	11.0	37.3	9.6	33.3	10.0	31.7	7.4	27.3	9.8	33.9
LLaVA-v1.6-mistral-7b	10.9	25.7	8.8	22.0	9.9	24.0	10.4	26.7	10.1	25.7	6.0	19.3	9.4	23.9
InstructBLIP	13.7	45.0	9.4	36.7	12.0	37.3	8.9	31.7	10.2	35.7	6.6	25.0	10.1	35.2
MiniGPT-4	12.0	38.7	9.8	33.0	11.3	34.7	10.6	32.7	11.3	33.3	8.8	31.0	10.6	33.9
Tinyllava	13.6	34.0	8.1	20.3	9.3	22.0	10.4	24.0	9.2	22.3	6.5	19.0	9.5	23.6
Phi3V	11.3	27.3	9.8	25.7	11.3	20.7	9.6	26.7	9.8	25.7	6.8	22.3	9.8	24.7
Fuyu	17.7	60.3	17.1	60.7	24.7	61.3	17.7	55.0	19.8	57.7	10.8	48.7	18.0	57.3
Idefics2-8b	9.5	23.7	8.6	23.3	9.9	24.7	11.4	31.7	9.2	26.0	6.5	20.0	9.2	24.9
mPLUG-Owl2	13.5	40.7	11.6	37	17.2	44.7	13.9	38.7	10.4	35	7.9	27.3	12.4	37.2
Qwen-VL	13.2	42.7	8.3	30.3	14.8	39.3	10.2	30.7	10.2	35.0	5.0	19.7	10.3	33.0

Table 1: Style-Chair Benchmark with prompt: *"Please describe this image in detail"*. Lower indicate fewer hallucinations.

Visual Sensitivity with Style-Aware Score (SAS)

To quantify the influence of visual style on token generation, we propose *Style-Aware Score* (SAS), which measures the alignment between intermediate visual representations and the final output tokens. Specifically, SAS captures the contribution of early-layer visual embeddings to token prediction by analyzing the intermediate hidden states within the transformer decoder. Given a set of candidate tokens selected from the final decoder layer's logits, we compute the SAS by aggregating token logits across visual token positions and selected intermediate layers. Let $\mathcal{L} \subset \{1, 2, \ldots, N-1\}$ denote a predefined set of candidate transformer layers (e.g., early or deep layers). For each selected layers $l \in \mathcal{L}$, we extract logits $\mathbf{Z}_l \in \mathbb{R}^{T \times |V|}$ by projecting the hidden states h^l via the output embedding matrix W_U, i.e., $\mathbf{Z}_l = W_U h^l$. We then isolate the logits corresponding to the visual token positions, yielding $\mathbf{Z}_l^{(v)} \in \mathbb{R}^{P \times |V|}$. The Style-Aware Score for a candidate token c at layer l is defined as:

$$\text{SAS}_l(c) = \sum_{p=1}^{P} softmax(\mathbf{Z}_l^{(v)}(c)) \quad (2)$$

Where P is the visual token length. This score quantifies the model's reliance on visual features during token generation. **A higher SAS implies stronger alignment between visual input and the generated token, indicating stronger grounding.** SAS provides a way to quantitatively assess a model's sensitivity to visual inputs.

Reduced Style Awareness Produces Hallucination

Prior studies have shown that language model priors often dominate visual inputs, with attention skewed toward object-like tokens (Wang et al. 2024a; Jiang et al. 2024). Motivated by these findings, we hypothesize that intermediate representations in early decoder layers encode varying degrees of visual awareness, which directly influences hallucination. Fig. 1 illustrates this hypothesis: tokens grounded in real visual objects (e.g., "tree") exhibit higher visual awareness, while hallucinated tokens show weaker spatial alignment. To empirically validate this observation, we conduct a layer-wise analysis of SAS trajectories throughout the transformer stack using the prompt "Please describe this image in detail". In the Style-POPE benchmark, each image is paired with six object existence questions (e.g.,"Is there a bottle

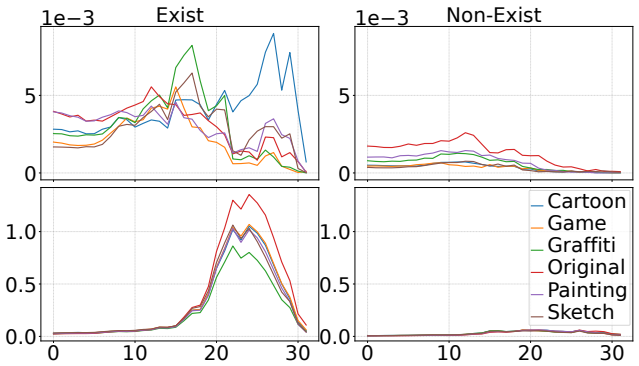

Figure 4: Style-awareness statistics across layer (Horizontal axis) using MiniGPT-4 (above) and InstructBlip.

Algorithm 1: Style-Aware Visual Early Revision (SAVER)

Require Input IDs $\mathbf{x}$, beam width B, early layers $\mathcal{L}$, thresholds k, p, scale α

Initialize beams $\{\mathbf{x}^{(b)}\}_{b=1}^{B}$

while condition **do**
  Decode: Forward decoder to get final-layer logits $\mathbf{z}_t^N$ and early-layer logits $\{\mathbf{z}_t^l\}_{l\in\mathcal{L}}$
  if first decoding step **then**
    Precompute SAS: cache $\{\mathrm{SAS}_l(c)\}_{l\in\mathcal{L},\, c\in V}$
  end if
  Candidate Filtering: $\mathcal{C}_t \leftarrow \textsc{TopK/TopP}(\mathbf{z}_t^N, k, p)$
  Style-Confidence Layer Selection:
    $\gamma = \max_{l\in\mathcal{L},\, c\in\mathcal{C}_t} \mathrm{SAS}_l(c)$
    $l^\star = \arg\max_{l\in\mathcal{L}} \max_{c\in\mathcal{C}_t} \mathrm{SAS}_l(c)$
  Logit Revision: $\hat{\mathbf{z}}_t = \mathbf{z}_t^N + \alpha\cdot\gamma\cdot(\mathbf{z}_t^{l^\star} \odot \mathbf{m}_t)$, where $\mathbf{m}_t$ masks $\mathcal{C}_t$
  Beam Update: Update beams using revised logits $\hat{\mathbf{z}}_t$
end while
return best completed beam(s)

in the image?"). A "no" label indicates the object is not present. For each question extract object names and their existence, and compute the average SAS values based on the predicted answer and the queried object. As shown in Fig. 4, tokens corresponding to real objects consistently exhibit significantly higher SAS scores than those associated with non-existent (hallucinated) objects, especially in the early layers. We further observe that SAS distributions are style-dependent, with peaks often occurring in earlier layers of the model.

Our Method

As discussed in the previous section, object hallucination in stylized images arises from the model's neglect of the corresponding visual patterns during decoding. To mitigate this issue, we propose **S**tyle-**A**ware **V**isual **E**arly **R**evision (**SAVER**), a test-time strategy that adjusts final predictions using early-layer representations. Notably, SAVER introduces no additional learnable parameters and can be seamlessly integrated into existing LVLMs and mitigate hallucination problems. SAVER operates during decoding and consists of three key steps: (1) identifying a candidate set of plausible tokens using top-p filtering; (2) selecting the most visually grounded layer via the Style-Aware Score (SAS); and (3) revising the logits to better reflect visual evidence from that layer. The algorithm pipeline is presented in **Algorithm 1**, which outlines the step-by-step implementation of SAVER at test time.

Style-Aware Layer Selection. Our previous analysis shows that correct tokens are highly correlated with the activation regions in early layers where visual features dominate. In this vein, we define a candidate set of transformer layers $\mathcal{L} \subset \{1, \ldots, N-1\}$ and identify a candidate token set $\mathcal{C}_t$ by applying top-p filtering to the final-layer logits $\mathbf{z}_t^N \in \mathbb{R}^{|V|}$ and top-k for the tokens. For each layer $l \in \mathcal{L}$, we compute a style-confidence score based on the maximum Style-Aware Score (SAS; see Eq. (2)) among the candidate tokens:

$$l^\star = \arg\max_{l\in\mathcal{L}} \sigma_l, \quad \gamma = \sigma_{l^\star}, \quad \text{where} \quad \sigma_l = \max_{c\in\mathcal{C}_t} \mathrm{SAS}_l(c). \tag{3}$$

Here, $\gamma \in [0, 1]$ quantifies the influence of stylistic features at the selected layer $l^\star$ and adaptively modulates visually relevant outputs. This mechanism allows SAVER to dynamically identify which layer provides the most relevant visual grounding at each decoding step, balancing between overfitting to style and ignoring visual context. The use of maximum SAS ensures that even if a single token strongly activates visual features at a certain layer, that signal is preserved in layer selection. In practice, we find that this adaptive grounding depth is critical for generalization across diverse styles and model architectures.

Logit Revision. SAVER refines the final-layer logits $\mathbf{z}_t^N$ by incorporating evidence from the selected style-sensitive layer $l^\star$. Let $\mathbf{m}_t \in \{0, 1\}^{|V|}$ denote a binary mask that activates only the candidate tokens in $\mathcal{C}_t$, which suppresses noise and stabilizes the logit revision. The revised logits $\hat{\mathbf{z}}_t$ are computed as:

$$\hat{\mathbf{z}}_t = \mathbf{z}_t^N + \alpha \cdot \gamma \cdot \left(\mathbf{z}_t^{l^\star} \odot \mathbf{m}_t\right), \tag{4}$$

where α is a scalar hyperparameter and $\odot$ is element-wise multiplication. This formulation selectively amplifies predictions that are grounded in style-aware visual evidence, while suppressing tokens that may arise solely due to stylistic noise. By scaling with both α and the layer-specific confidence γ, the method modulates its correction strength based on how strongly the model attends to visual input at $l^\star$. Since SAVER operates per decoding step, it enables fine-grained correction without modifying the backbone model or compromising generation fluency. This makes it applicable in real-world scenarios with diverse data distributions and effective even under domain shifts.

Experiments

Implementation Details

Evaluation Models and Settings. We evaluate our proposed method and baseline approaches on four representative LVLMs: InstructBLIP (Dai et al. 2023), MiniGPT-4 (Zhu et al. 2023), and LLaVA-1.5 (Liu et al. 2023a),

Model	Method	Cartoon		Game		Graffiti		Painting		Sketch		Original		Average	
		Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs	Ci	Cs
InstructBLIP	Greedy	13.7	45.0	9.4	36.7	12.0	37.3	8.9	31.7	10.2	35.7	6.6	25.0	10.1	35.2
	Beam	11.8	41.3	9.7	34.0	10.8	34.7	7.8	27.3	10.4	38.3	7.1	28.7	9.6	34.1
	Dola	18.8	42.7	19.3	48.3	20.2	46.3	17.8	43.7	17.6	43.7	12.7	39.0	17.7	44.0
	OPERA	12.4	38.3	10.7	35.0	12.2	37.3	8.8	31.0	9.9	34.7	7.0	25.7	10.2	33.7
	Deco	10.9	36.3	8.9	**28.0**	10.6	30.7	9.4	26.0	8.2	**25.3**	5.5	22.3	8.9	28.1
	AGLA	12.4	41.0	9.6	36.0	12.8	38.0	9.9	33.0	10.4	37.7	7.0	27.3	10.4	35.5
	SAVER(Ours)	**9.9**	**32.0**	**8.0**	30.0	**9.9**	**28.0**	**7.4**	**25.3**	**7.8**	27.0	**5.3**	**21.3**	**8.1**	**27.3**
LLaVA-1.5	Greedy	12.0	43.3	9.1	31.7	**11.4**	37.7	11.2	34.7	11.4	37.7	6.8	26.7	10.3	35.3
	Beam	11.5	39.3	9.3	32.3	11.9	33.3	10.2	34.3	10.5	36.7	6.0	24.0	9.9	33.3
	Dola	15.4	44.0	14.1	43.0	16.9	49.0	14.7	42.0	14.9	46.7	10.6	35.7	14.4	43.4
	OPERA	12.1	38.7	9.6	33.0	12.4	35.7	10.1	31.7	10.7	31.7	**5.8**	24.7	10.1	32.6
	Deco	**11.2**	**31.3**	8.9	30.0	11.9	32.7	9.4	31.3	**9.4**	**30.7**	5.9	**21.3**	**9.5**	**29.6**
	AGLA	11.3	38.0	9.3	31.3	12.5	40.0	10.4	33.3	11.9	39.3	6.7	26.7	10.4	34.8
	SAVER(Ours)	**11.2**	32.7	**8.8**	**29.7**	**11.4**	**30.3**	**9.3**	**29.0**	10.6	32.0	6.6	26.0	9.7	30.0
LLaVA-1.6	Greedy	12.3	27.0	8.8	22.0	9.9	24.0	10.4	26.7	10.1	25.7	6.0	19.3	9.4	23.9
	Beam	11.9	29.0	9.6	24.3	10.1	22.7	10.7	26.7	9.3	25.3	5.9	18.0	9.6	24.3
	Dola	13.6	31.0	8.1	18.3	10.2	**17.7**	8.9	**18.7**	8.7	20.0	**5.1**	**12.7**	9.1	**19.7**
	OPERA	**10.1**	**23.0**	8.6	21.5	12.1	27.1	9.8	25.6	9.2	24.7	6.0	17.4	9.3	23.2
	Deco	12.8	30.3	9.7	24.3	10.9	25.3	**8.4**	22.0	9.3	26.7	6.1	18.7	9.5	24.6
	AGLA	10.8	28.0	9.2	23.3	9.9	24.0	10.5	27.3	8.9	25.7	6.0	20.3	9.2	24.8
	SAVER(Ours)	13.0	31.7	**7.4**	**16.3**	**9.7**	20.3	9.0	19.7	**8.3**	**18.7**	5.2	13.0	**8.8**	20.0
MiniGPT-4	Greedy	12.0	38.7	9.8	33.0	11.3	34.7	10.6	32.7	11.3	33.3	8.8	31.0	10.6	33.9
	Beam	11.5	34.3	10.7	34.3	9.5	30.0	10.0	33.7	11.2	34.3	8.1	28.7	10.2	32.6
	Dola	16.4	40.3	12.3	34.3	15.6	36.7	13.4	35.0	13.5	36.0	9.8	28.7	13.5	35.2
	OPERA	11.5	34.7	10.9	34.7	10.4	31.0	9.6	31.0	10.7	34.3	8.2	29.0	10.2	32.5
	Deco	11.3	38.3	9.2	29.3	10.2	28.3	8.6	27.7	9.8	**28.0**	7.0	24.7	9.4	29.4
	AGLA	13.2	43.7	11.2	43.2	13.5	45.3	11.5	39.0	12.9	42.0	10.7	42.3	12.2	42.6
	SAVER(Ours)	**9.4**	**30.0**	**8.8**	**28.7**	**9.3**	**28.0**	**8.5**	**25.3**	**8.1**	**28.0**	**5.8**	**24.0**	**8.3**	**27.3**

Table 2: CHAIR hallucination evaluation results (lower is better). Best values are bolded; second-best are underlined.

LLaVA-1.6 (Liu et al. 2024). The maximum length of the generated sequence is set to 64 tokens, with a repetition penalty of 1.0. Unless otherwise specified, decoding is performed using a fixed temperature of 0 and top-k=1. For beam search, we adopt a beam width of 3 while maintaining the same temperature setting.

Baselines. We compare SAVER with two baseline decoding strategies (greedy decoding and beam search) as well as three SOTA hallucination mitigation methods, detailed as follows: Dola (Chuang et al. 2023) is specifically designed for alleviating hallucinations in factual tasks for LLMs by reducing shallow semantic influences to improve the factuality of the final layer's output. OPERA (Huang et al. 2024) dynamically penalizes overconfident tokens based on the emergence of aggregation patterns, while proposing a retrospective allocation strategy to avoid cases where hallucinations have already occurred. Deco (Wang et al. 2024a) adaptively chooses relevant layers and integrates their knowledge into the final layer to adjust outputs. AGLA (An et al. 2024) uses an ensemble of global features for response generation and local features to mitigate hallucination. For all baseline methods, we use their official implementations and follow the recommended hyperparameter settings from the released source code to ensure fair comparisons.

Benchmark and Metrics. We evaluate the effectiveness, generalizability, and captioning quality of our method across five challenging benchmarks in both stylized and real-world scenarios: • **CHAIR.** Using the prompt "*Please describe the image in detail*," we assess hallucination rates with CHAIRi and CHAIRs on our constructed dataset. Additionally, we evaluate captioning quality using BLEU-1/2/3/4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), and ROUGE-L (Lin 2004) which can be found in Appendix. • **POPE.** Following the official POPE protocol, we report F1 scores as the primary metric. In the Appendix, we further incorporate ACC, Precision, and Recall to comprehensively evaluate our method. • **MME** (Fu et al. 2024) is a practical benchmark encompassing 14 sub-tasks, including OCR, visual knowledge, object recognition, and relational reasoning. • **Real-World Cases.** To evaluate performance beyond stylized images, we construct a dataset containing depth, thermal, medical, and RGB images with carefully designed query prompts. Additional details are available in the Appendix. • **AMBER.** (Wang et al. 2023a) To evaluate hallucination in more dimisions such as attribute, relation, and existence, we conduct experiments on the AMBER benchmark. More details are provided in the Appendix.

Method	Adversarial			Popular			Random			Overall Avg		
	M.	L.	I.	M.	L.	I.	M.	L.	I.	M.	L.	Ins.
Dola	26.1	63.9	62.9	26.4	64.3	63.2	25.8	64.2	64.8	26.1	64.1	63.7
Deco	67.0	73.3	**73.6**	66.8	74.6	72.2	70.8	80.5	82.7	68.2	76.2	76.2
Ours	**67.6**	**75.3**	73.5	**67.4**	**77.6**	**72.7**	**73.6**	**84.4**	**84.2**	**69.5**	**79.1**	**76.8**

Table 3: Results on Style-POPE hallucination. "M"., "L." and "I." stand for MiniGPT-4, LLaVA-1.5, and InstructBLIP.

Experimental Results

Experiments on Style-CHAIR. As shown in Tab. 2, our method consistently reduces hallucinations across all styles, achieving the lowest average CHAIR across all evaluated models. Notably, SAVER achieves the best over-

Model	Method	Scene	Num.	Text.	Code.	Total Score
LLaVA-1.5	Dola	58.6	60.0	82.5	**57.5**	258.6
	Deco	85.7	70.0	50.0	47.5	253.2
	Ours	**88.6**	**77.5**	**115.0**	45.0	**326.1**
InstructBLIP	Dola	81.4	**55.0**	57.5	45.0	238.9
	Deco	81.4	45.0	**65.0**	**72.5**	263.9
	Ours	**107.1**	50.0	57.5	55.0	**269.6**

Table 4: Results on MME recognition related to the subtasks of commonsense reasoning, numerical calculation, text translation, and code reasoning.

Method	LLaVA-1.5		MiniGPT-4		InstructBLIP	
	ACC	F1	ACC	F1	ACC	F1
Dola	62.2	62.4	**55.4**	37.2	**65.8**	43.8
Deco	57.4	69.4	50.2	**66.4**	60.0	68.2
Ours	**64.0**	**71.8**	51.4	**66.4**	65.0	**68.6**

Table 5: Average Results on Real-World Benchmarks.

all performance on MiniGPT-4 and InstructBLIP, clearly outperforming the SOTA method Deco. On LLaVA-1.5, SAVER achieves competitive results, closely matching the best scores. These results demonstrate the effectiveness of SAVER in mitigating object hallucinations for stylized images across diverse models and styles. Since captioning quality is critical for real-world applications, we further report the performance of different mitigation methods in terms of captioning quality in Tabs. 10, 11, and 12 in the Appendix. Compared to existing baselines, SAVER consistently exhibits outstanding text captioning performance.

Experiments on Style-POPE and MME Benchmarks. As shown in Tab. 3, SAVER consistently outperforms prior methods under various configurations on the Style-POPE benchmark, demonstrating superior robustness in mitigating hallucinations across adversarial, popular, and random settings. The detailed results are shown in Tab. 13, Tab. 14, and Tab. 15 in the Appendix. Additionally, in Tab. 4 and Tab. 16 (Appendix), SAVER achieves the highest scores across both perception and recognition tasks on the challenging MME benchmark, significantly improving performance for LLaVA-1.5 and InstructBLIP. Beyond stylized image captioning, SAVER exhibits strong generalizability to mitigate hallucinations in various practical challenging tasks.

Experiments on Real-World Benchmarks. To further examine SAVER's practicability, we evaluate it on real-world scenarios, including depth, thermal, and medical images. As the average results shown in Tab. 5, SAVER consistently achieves the highest performance in different modalities. Specifically, it obtains the best average F1 scores of 71.8% on LLaVA-1.5 and 68.6% on InstructBLIP. These results further validate the effectiveness of SAVER in real-world scenarios. The detailed results and discussions can be found in the Appendix.

Experiments on AMBER Benchmarks. Previous experiments comprehensively demonstrated SAVER's effectiveness in mitigating object hallucination. Herein, we further evaluate SAVER's generalizability on the AMBER dataset, which consists of three types of hallucinations, including existence, attribute, and relation. AMBER comprises 1,004 images, each paired with the corresponding designed questions and annotated labels. As shown in Tab. 6, SAVER consistently achieves the best or second-best Acc/F1 scores compared to the SOTA methods, exhibiting its strong generalizability to various hallucination types. This can be attributed to our visual early revision design that drives the model to pay more attention to visual signals.

Method	Existence		Attribute		Relation	
	Acc	F1	Acc	F1	Acc	F1
Dola	66.0	79.5	48.1	44.7	22.1	25.2
Deco	63.1	62.5	67.2	53.5	**69.5**	48.1
Ours	**67.7**	**80.7**	**68.3**	**58.2**	69.1	**62.5**

Table 6: LLaVA-1.5-7b results on AMBER.

Ablation Study

Tabs. 18, 19, 20 and Fig. 6 in the Appendix detail the ablation experimental results. We vary five components—scale factor α, confidence threshold p, candidate set size k, number of image-representative tokens N_i, and early-exit depth—to quantify their contributions to hallucination mitigation and caption fluency. For scale factor, $\alpha = 0.6$ consistently balances visual grounding and language quality, attaining the lowest hallucination scores on LLaVA-1.5. For token filtering, higher thresholds reduce low confidence tokens. And $p = 0.9$ achieves the strongest average results for MiniGPT-4 and InstructBLIP, while lower p preserves diversity in challenging styles at the cost of stability. In the candidate size, $k = 20$ emerges as a robust optimum across models, and larger k introduces spurious evidence. Visual awareness is best supported by moderate N_i (50–100), as larger values can introduce low-confidence tokens that increase hallucination risks. Finally, the choice of early exit depth can greatly affect hallucinations. Experiments show that "Standard" provides the most stable performance.

Conclusion

This work studies the high hallucination risk of existing LVLMs when understanding stylized images and how to mitigate it. By constructing a stylized dataset and a comprehensive benchmark, we demonstrate that stylized inputs significantly increase hallucination rates. To explore the underlying causes, we analyze the correlation between generated tokens and image tokens, revealing that the later layers of LVLMs tend to suppress visual information and rely more heavily on language priors. To address this, we propose a training-free mitigation method, SAVER, which dynamically corrects generated tokens by retrieving optimal early layers with dense visual correlation. Extensive experiments across diverse models, datasets, and tasks validate the effectiveness of SAVER. We hope this work contributes to the development of more trustworthy LVLMs and facilitates their application in challenging scenarios.

Acknowledgments

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab, School of Electrical & Electronic Engineering, Nanyang Technological University (NTU), Singapore. This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

References

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI, G. 2024. Gemini: A Large Language Model.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

An, W.; Tian, F.; Leng, S.; Nie, J.; Lin, H.; Wang, Q.; Dai, G.; Chen, P.; and Lu, S. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bavishi, R.; Elsen, E.; Hawthorne, C.; Nye, M.; Odena, A.; Somani, A.; and Taşırlar, S. 2023. Introducing our Multimodal Models.

Brie, P.; Burny, N.; Sluÿters, A.; and Vanderdonckt, J. 2023. Evaluating a large language model on searching for gui layouts. *Proceedings of the ACM on Human-Computer Interaction*, 7(EICS): 1–37.

Cai, R.; Song, Z.; Guan, D.; Chen, Z.; Li, Y.; Luo, X.; Yi, C.; and Kot, A. 2024. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. In *European Conference on Computer Vision*, 340–358. Springer.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chen, L.; Sinavski, O.; Hünermann, J.; Karnsund, A.; Willmott, A. J.; Birch, D.; Maund, D.; and Shotton, J. 2024a. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 14093–14100. IEEE.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024c. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Chiang, W.-L.; and Li, Z. 2 May 2025. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; Wu, Y.; and Ji, R. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394.

Hu, M.; Pan, S.; Li, Y.; and Yang, X. 2023. Advancing medical imaging with language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13418–13427.

Hwang, S.; Park, J.; Kim, N.; Choi, Y.; and Kweon, I. S. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiang, Z.; Chen, J.; Zhu, B.; Luo, T.; Shen, Y.; and Yang, X. 2024. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*.

Jing, L.; and Du, X. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

Jing, L.; Li, R.; Chen, Y.; and Du, X. 2023. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Zhu, Y.; Kato, K.; Kondo, I.; Aoyama, T.; and Hasegawa, Y. 2023b. Llm-based human-robot collaboration framework for manipulation tasks. *arXiv preprint arXiv:2308.14972*.

OpenAI. 2023. GPT-4 Technical Report. https://openai.com/research/gpt-4. Accessed: 2024-10-18.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 746–760. Springer.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, C.; Chen, X.; Zhang, N.; Tian, B.; Xu, H.; Deng, S.; and Chen, H. 2024a. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*.

Wang, H.; Wang, Q.; Bai, X.; Qin, Z.; and Chen, A. 2024b. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation. *arXiv preprint arXiv:2404.02733*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; et al. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Wang, L.; He, J.; Li, S.; Liu, N.; and Lim, E.-P. 2024c. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, 32–45. Springer.

Wang, S.; Zhao, Z.; Ouyang, X.; Wang, Q.; and Shen, D. 2023b. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

Xie, Y.; Li, G.; Xu, X.; and Kan, M.-Y. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*.

Xing, Y.; Li, Y.; Laptev, I.; and Lu, S. 2024. Mitigating object hallucination via concentric causal attention. *Advances in Neural Information Processing Systems*, 37: 92012–92035.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13040–13051.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024a. A survey on multimodal large language models. *National Science Review*, 11(12).

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2024b. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12): 220105.

Zhang, J.; Wang, T.; Zhang, H.; Lu, P.; and Zheng, F. 2024. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, 196–213. Springer.

Zhao, Z.; Tang, J.; Wu, B.; Lin, C.; Wei, S.; Liu, H.; Tan, X.; Zhang, Z.; Huang, C.; and Xie, Y. 2024. Harmonizing visual text comprehension and generation. *arXiv preprint arXiv:2407.16364*.

Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; and Huang, L. 2024. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. arXiv:2402.14289.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

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