arXiv:2411.14279v1 [cs.CV] 21 Nov 2024

Looking Beyond Text: Reducing Language bias in Large Vision-Language Models via Multimodal Dual-Attention and Soft-Image Guidance

Haozhe Zhao[♠*], Shuzheng Si[♢*], Liang Chen[♠], Yichi Zhang[♠]
Maosong Sun[♢], Mingjia Zhang[♡†], and Baobao Chang[♠†]
[♠]Peking University [♡]University of Illinois Urbana-Champaign [♢]Tsinghua University

Abstract

*Large vision-language models (LVLMs) have achieved impressive results in various vision-language tasks. However, despite showing promising performance, LVLMs suffer from hallucinations caused by language bias, leading to diminished focus on images and ineffective visual comprehension. We identify two primary reasons for this bias: 1). Different scales of training data between the pretraining stage of LLM and multimodal alignment stage. 2). The learned inference bias due to short-term dependency of text data. Therefore, we propose LACING, a systemic framework designed to address the language bias of LVLMs with Mu**L**timodal Du**A**l-attention Me**C**han**I**sm (MDA) a**N**d Soft-Image **G**uidance (SIG). Specifically, MDA introduces a **parallel dual-attention mechanism** that enhances the integration of visual inputs across the model. SIG introduces a **learnable soft visual prompt** during training and inference to replace visual inputs, designed to compel LVLMs to prioritize text inputs. Then, SIG further proposes a novel decoding strategy using the soft visual prompt to mitigate the model's over-reliance on adjacent text inputs. Comprehensive experiments demonstrate that our method effectively debiases LVLMs from their language bias, enhancing visual comprehension and reducing hallucinations without requiring additional training resources or data. The code and model are available in `lacing-lvlm.github.io`.*

1. Introduction

The emergence of Large Language Models (LLMs) [10, 42, 56, 57] has marked a significant milestone across various natural language processing tasks [40, 62]. Leveraging the capabilities of LLMs, the development of Large Vision-Language Models (LVLMs) [38, 41, 43, 55] has been accelerated through the integration of pre-trained visual encoders into these LLMs [3, 7, 26, 29]. This integration allows LVLMs to process both visual and text inputs, and further enables various potential applications such as autonomous driving [36, 52, 60] and medical assistants [24].

Despite their impressive capabilities in handling both two modalities, state-of-the-art LVLMs often produce erroneous or hallucinatory responses irrelevant to the input images [5, 13, 21]. The main reason behind the hallucinations in LVLMs is referred to as language bias [67], i.e., these models sometimes "ignore" the visual input and generate the text response solely on the text inputs.

However, previous studies have not yet thoroughly investigated how this language bias originates. We suggest that this bias potentially emerges for the following two reasons:

1. Different training scales between the LLM pre-training stage and LVLM alignment stage: The LLM backbone within LVLMs is pre-trained on the large-scale text corpus. Comparatively, the multimodal alignment stage in current SOTA LVLMs utilizes much fewer training samples and shorter training periods. For instance, Llama3 [10] undergoes pre-training with 15T tokens, whereas the multimodal alignment training for LLaVA [23, 29–31] utilizes about 558k to 1.3 million data. The gap between the training scales of the two stages inevitably makes the pretraining distribution dominate the generation of LVLMs [44], leading to difficulties for LVLMs to fully utilize the visual inputs, e.g., LVLMs often allocate a low attention weight to the visual inputs in more than 90% layers [6, 66].

2. The learned inference bias due to the short-term dependency of text data: Intuitively, a word in a text sequence exhibits a stronger associative bond with adjacent words than those further apart [2, 9, 61], i.e., the short-term dependency of text data. LLMs pre-trained on large-scale text corpora can easily capture and memorize such short-term dependency. Typically, these models assign higher attention weights to adjacent tokens when processing the text data. However, this learned pattern may not be suitable for multi-modal inputs. Visual inputs frequently appear at the beginning of the full inputs in current LVLMs, causing these models to anchor on the nearby text instructions and neglect to draw logical inferences from the actual visual inputs dur-

---

[*] Equal Contribution.
[†] Corresponding Author.

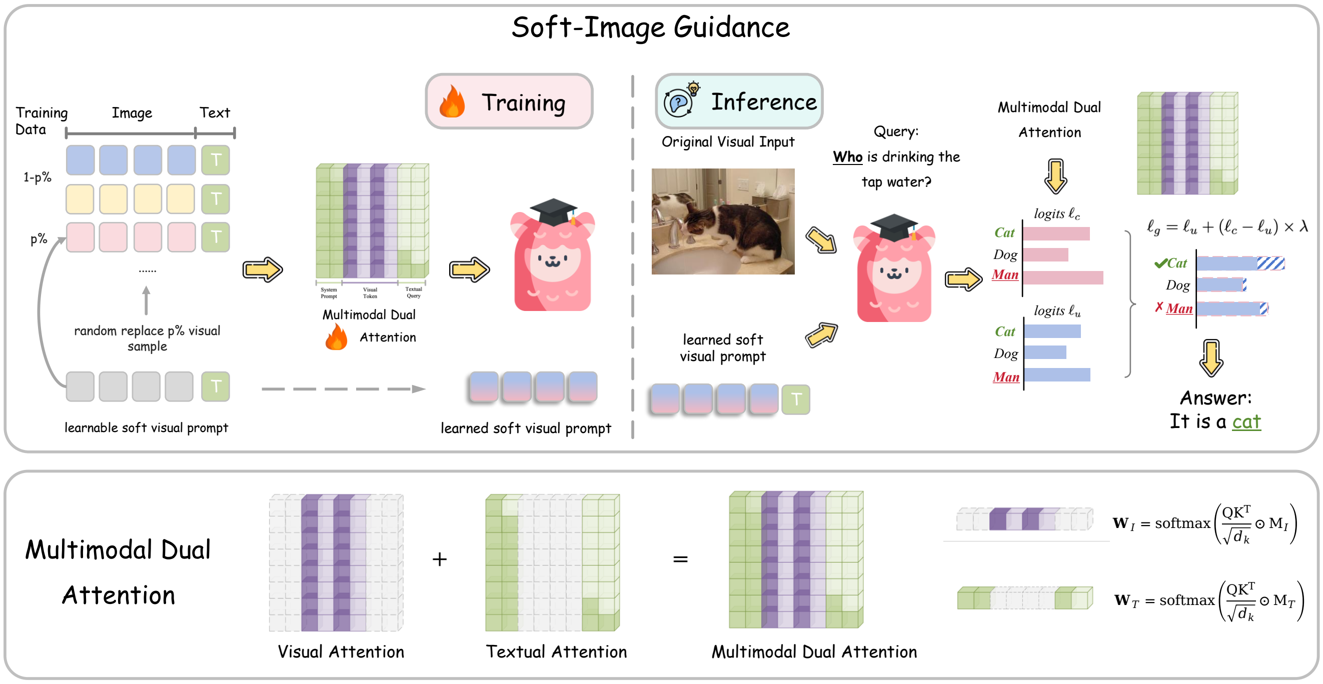

Soft-Image Guidance
Training Data
Image
Text
1-p%
p%
T
......
random replace p% visual sample
learnable soft visual prompt
Training
System Prompt
Visual Token
Textual Query
Multimodal Dual Attention
learned soft visual prompt
Inference
Original Visual Input
Query: Who is drinking the tap water?
learned soft visual prompt
Multimodal Dual Attention
logits ℓ_c
Cat
Dog
Man
logits ℓ_u
Cat
Dog
Man
$\ell_g = \ell_u + (\ell_c - \ell_u) \times \lambda$
✓Cat
Dog
✗ Man
Answer: It is a cat
Multimodal Dual Attention
+
=
Visual Attention
Textual Attention
Multimodal Dual Attention
$\mathbf{W}_I = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M_I\right)$
$\mathbf{W}_T = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M_T\right)$

Figure 1. Overview of **LACING**, a systemic framework, which consists of **Multimodal Dual Attention (bottom)** and **Soft-Image Guidance (above)** to mitigate language bias of LVLMs. MDA proposes a parallel dual-attention mechanism that constructs two separate attention for visual and text inputs during both training and inference. SIG implements a learnable soft visual prompt during training to replace visual inputs. This soft prompt serves to maintain input patterns while compelling models to prioritize text inputs during inference.

ing the text response generation.

These two reasons lead to a systemic bias in LVLMs, originating from both training and inference stages. Consequently, a critical question arises: ***How can we effectively mitigate the language bias inherent in LVLMs from both the training and inference perspectives?*** Therefore, we propose **LACING**, a systemic framework designed to address the language bias of LVLMs with MuLtimodal DuAl-attention MeChanIsm aNd Soft-Image Guidance.

Exploring methods to mitigate the dominant influence of the LLM's pretraining distribution [44], which leads to the neglect of visual inputs in most layers [6], remains an area that still requires further exploration. To address this issue, we propose the **Multimodal Dual-Attention Mechanism (MDA)**. MDA introduces a parallel dual-attention mechanism to effectively handle both modalities separately during the multimodal alignment stage. Specifically, MDA first independently calculates attention weights across various modalities, then fuses the two different weights to get the final attention map. It enforces the LVLM to assure considerable attention weights to visual inputs, ensuring integration of visual inputs across all layers. Furthermore, MDA retains causal attention for text inputs while allowing bidirectional attention to visual inputs. It helps LVLMs preserve the text comprehension ability of LLMs while effectively capturing the global structure of visual inputs, eventually focusing on visual inputs across all layers during training.

To mitigate the learned inference bias due to the short-term dependency of text data, we introduce **Soft-Image Guidance (SIG)**, designed to enhance the visual guidance during LVLMs' response generation. The core idea behind SIG is to effectively model the current LVLMs' over-reliance on text inputs, i.e., language bias, and reduce such over-reliance with a novel decoding strategy. Specifically, SIG first introduces a **learnable soft visual prompt** to replace the visual input and construct a multimodal-null input for the model. This learnable soft visual prompt serves as a placeholder to maintain the input pattern (e.g., the input length and modalities) and as an indicator, compelling the model to prioritize text inputs. Different from previous works [22, 66] attempting to model such over-reliance by removing visual inputs or adding random noise on input images, our designed soft visual prompt can maintain the consistent input pattern without introducing uncontrollable informational perturbations. To get the final soft visual prompt, we replace visual inputs with the soft visual prompt at a given probability during the multimodal alignment stage. The well-learned soft visual prompt is used to replace visual inputs in original inputs to suggest the model prioritize the text input, namely multimodal-null inputs. Finally, we contrast output distributions from original and multimodal-null inputs to get the final output, mitigating the

LVLM's language bias during the response generation. This adjustment ensures that each token in the model's responses accounts for visual input more critically.

Comprehensive experiments demonstrate the effectiveness of our method in mitigating language bias in LVLMs, thereby enhancing visual comprehension and reducing visual hallucination. We observe significant improvements, particularly in free-form generation and visual hallucinations reduction (e.g., 11.8-point boost on LLaVA-Bench [29] and a 40% improvement on Object Hallucinations Benchmark [48, 63]). Notably, our approach achieves consistent improvement without requiring additional training resources or data beyond standard LVLM multimodal alignment setups [30]. Our analysis further confirms the efficacy of MDA in enabling the LVLM to fully utilize the visual inputs, as well as the robustness of the SIG for reducing hallucinations and improving visual comprehension.

2. Related Work

2.1. Language Bias in LVLMs

Despite the impressive capabilities of LVLMs [11, 23, 38, 41, 43, 55, 58], these models still struggle with generating responses irrelevant to the input images [21, 32], e.g., falsely perceiving the presence of non-existent objects in images [70]. Zhao et al. [67] first identify this issue in LVLMs and name it as *language bias*, i.e., LVLMs often ignore visual inputs and solely rely on text inputs, leading to hallucinations. Chen et al. [5] also observes that LVLMs tend to solely use the textual knowledge from LLMs to answer the questions. Chen et al. [6] analyzes the attention distribution of state-of-the-art LVLMs, revealing an inefficient attention mechanism wherein attention computation over visual inputs is extremely inefficient in the deeper layers of LVLMs. Moreover, Zhang et al. [66] note that LVLMs tend to allocate more attention to text inputs and increasingly prioritize them as response generation progresses. These works indicate that LVLMs often allocate a low attention weight to the visual inputs, thus limiting the use of the given image. Consequently, solutions for mitigating such bias remain a crucial problem and means to further improve the performance of LVLMs. We argue that there are two main reasons that lead to the language bias in LVLMs, including (1) the different training scales between the pretraining stage and multimodal alignment stage, and (2) the learned inference bias due to the short-term dependency of text data. We further propose a systematic method LACING to address the language bias from both the training and inference perspectives

2.2. Addressing Language Bias in LVLMs

Given the language bias within current LVLMs, LVLMs exhibit similar hallucination problems as the textual LLMs [15]. Moreover, LVLMs show unique types of hallucination, such as object hallucination [25, 48], i.e., LVLMs tend to encompass content not supported by associated images. Leng et al. [22] claim that object hallucination is caused by the dominant influence of the LLM's pretraining distribution, i.e., such hallucination is a primary manifestation of language bias. Recently, many methods have been proposed to mitigate the hallucination in LVLMs. LRV [28] attempts to apply supervised fine-tuning on a well-designed visual preference dataset to address this issue. LLaVA-BPO [44] further proposes the pipeline to gather preference datasets and conduct preference learning to mitigate this type of hallucination. Additionally, LLaVA-RLHF [54] and RLHF-V [63] introduce reinforcement learning from human feedback into the LVLM training pipeline to align outputs with human intentions, thereby reducing hallucinations. However, these methods typically necessitate a substantial number of additional training data and computational resources. Thus, many methods such as VCD [22], IBD [72], VDD [66], and ICD [59] introduce training-free methods that contrast model outputs generated from original inputs and inputs without visual inputs (or inputs with masked visual inputs) to reduce the influence of textual LLMs. Nonetheless, these methods may introduce unexpected information during inference that was not present during training. This inconsistency between the inputs at the training and inference stages can result in suboptimal performance of these methods. Different from the two types of methods, our method has two notable strengths: (1) We propose our new method in a systematic way, addressing the language bias from both the training and inference perspectives without requiring additional training data; (2) Different from these methods only consider a part influence of language bias, e.g., object hallucination, our method addresses the language bias in two perspectives that further improve the performance of LVLMS in more general tasks.

3. Method

In this section, we present our approach for mitigating the language bias inherent in LVLMs using two complementary methods: the Multimodal Dual-Attention Mechanism and Soft-Image Guidance. These method address the language bias of LVLMs from both training and inference perspectives, each targeting distinct issues.

3.1. Multimodal Dual-Attention Mechanism

Most LVLMs project bidirectional visual inputs into the unidirectional LLM space using a relatively small amount of multimodal data [23, 29, 30, 71] compared to vast data scales employed in pretraining LLMs [10, 42]. These models treat visual inputs as a different form of text inputs in an autoregressive manner. The mismatch in both modeling and training scale leads LVLMs to partially adapt to

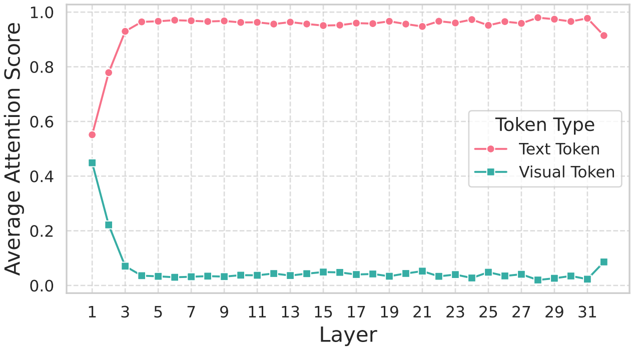


Figure 2. Average attention scores for output tokens towards text and visual tokens across different layers in LLaVA-1.5 [30], showing that only the first two layers apply considerable attention to visual tokens. In contrast, deeper layers largely neglect them.

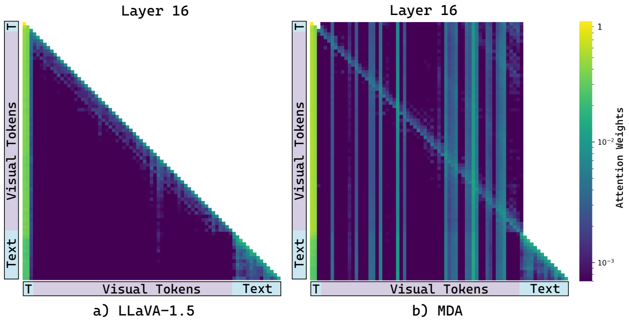
data distribution changes using only shallow layers during training with limited data [66]. Consequently, the model remains dominated by the LLM's pretraining distribution and lacks effective attention to visual inputs in deeper layers. As shown in Figure 2, LVLMs exhibit considerable attention toward visual inputs only in the first two layers [6], while deeper layers retain their original distributions, causing deeper layers of LVLMs to ignore visual inputs.

To address this issue, we propose **Multimodal Dual-Attention Mechanism (MDA)**, which introduces a parallel dual-attention mechanism that preserves separate attention metrics for visual and text inputs in the LVLMs. It enforces the LLM to allocate sufficient attention toward visual inputs and encourages LVLMs to fully leverage their LLM backbone for visual comprehension during training. Specifically, MDA is parallel dual-attention strategy that retain unidirectional, causal attention for text inputs while allowing bidirectional attention towards visual inputs for comprehensive understanding. As illustrated in Equation 1, Given the multimodal input $\mathbf{S} = \langle s_1, s_2, \ldots, s_N \rangle$, s_n means the token in the input. To independently calculate attention weights across two modalities, we define two attention masks: visual token mask $\mathbf{M}_{\mathcal{I}}$ for visual tokens $\mathcal{I}$ and text token mask $\mathbf{M}_{\mathcal{T}}$ for text tokens $\mathcal{T}$:

$$\begin{aligned} \mathbf{M}_{\mathcal{I}}[i,j] &= \begin{cases} 1, & \text{if } s_j \in \mathcal{I}, \\ 0, & \text{otherwise}, \end{cases} \\ \mathbf{M}_{\mathcal{T}}[i,j] &= \begin{cases} 1, & \text{if } s_j \in \mathcal{T} \ \ \& \ \ i \leq j, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{1}$$

We can use designed attention masks to calculate attention weights of visual($\mathbf{W}_{\mathcal{I}}$) and test tokens($\mathbf{W}_{\mathcal{T}}$):

$$\begin{aligned} \mathbf{W}_{\mathcal{I}} &= \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k} \odot \mathbf{M}_{\mathcal{I}}\right), \\ \mathbf{W}_{\mathcal{T}} &= \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{d_k} \odot \mathbf{M}_{\mathcal{T}}\right), \end{aligned} \tag{2}$$



a) LLaVA-1.5

b) MDA

Figure 3. Comparison of attention allocation between a standard LVLM (LLaVA-1.5) and our model trained with the Multimodal Dual-Attention (MDA) mechanism. Text tokens and visual tokens are indicated in blue and purple, respectively, in the sidebar.

where the $\mathbf{Q}, \mathbf{K}$ represent the query, key and in the self-attention mechanism of LVLMs.

Finally, the two attention weights, denoted as $(\mathbf{W}_{\mathcal{I}})$ and $(\mathbf{W}_{\mathcal{T}})$, are integrated and subsequently multiplied by $\mathbf{V}$, the value in the attention mechanism of LVLMs, to derive the final attention score $\mathbf{A}$ based on MDA.

$$\mathbf{A} = (\mathbf{W}_{\mathcal{I}} + \mathbf{W}_{\mathcal{T}})\,\mathbf{V}. \tag{3}$$

Parallel computation of attention weights guarantees each token separately receives attention from both visual and text inputs, balancing their contributions. It allows visual inputs to remain relevance across all layers, avoiding shallow adaptation and language bias. The dual-attention enhances the visual understanding of LVLMs while preserving its language modeling capabilities. MDA ensures that visual information is processed with bidirectional attention to capture spatial coherence, while text tokens continue to follow autoregressive patterns, critical for maintaining coherent language generation, as shown in Figure 3.

3.2. Soft-Image Guidance

A significant challenge in visual comprehension with LVLMs stems from their inference bias due to the short-term dependency of text data. Due to the sequential nature of language modeling—which prioritizes coherence and continuity—LVLMs tend to focus on nearby text tokens, often at the expense of the visual information that may be distant or disparate, as shown in Figure 4.

To mitigate the inference bias of LVLMs, we propose the Soft-Image Guidance (SIG), designed to enhance the guidance of visual inputs during LVLMs' response generation. To enhance the guidance of visual inputs in LVLMs, we formulate the visual comprehension mathematically. We consider the conditional probability $p(y_t \mid v)$ of generating a response token y_t given the visual input v. By applying

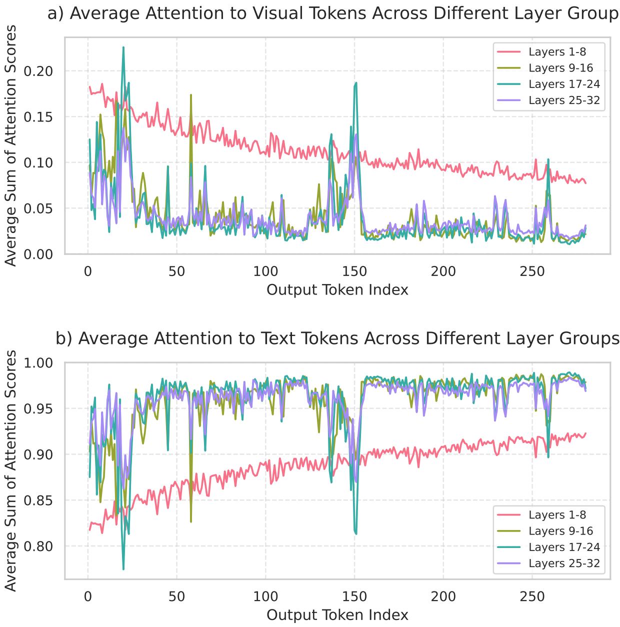


Figure 4. Attention allocation of LVLMs to visual and text tokens. Attention to visual tokens (a) decreases as response generation progresses, while attention to text tokens (b) increases.

Bayes' theorem, we have:

$$p(y_t \mid v) = \frac{p(v \mid y_t) \cdot p(y_t)}{p(v)} \tag{4}$$

Then we take the logarithm of both sides of Equation (4):

$$\log p(y_t \mid v) = \log p(v \mid y_t) + \log p(y_t) - \log p(v) \tag{5}$$

In Equation (5), $p(y_t)$ is the unconditional probability of generating the target token y_t without any visual input.

To amplify the influence of the visual input v on response generation, we introduce a scaling parameter λ for conditional probability $p(v \mid y_t)$. We adjust the $p(v \mid y_t)$ to obtain an enhanced version $\widehat{p}(y_t \mid v)$, defined as:

$$\log \widehat{p}(y_t \mid v) \propto \lambda \cdot \log p(v \mid y_t) + \log p(y_t) - \log p(v) \tag{6}$$

Our goal is to express $\log \widehat{p}(y_t \mid v)$ with known quantities. To this end, we expand $\log p(v \mid y_t)$ using Bayes' theorem:

$$\log p(v \mid y_t) = \log p(y_t \mid v) + \log p(v) - \log p(y_t) \tag{7}$$

Substituting Equation (7) into Equation (6), we obtain:

$$\begin{aligned}\log \widehat{p}(y_t \mid v) \propto \lambda \left(\log p(y_t \mid v) + \log p(v) - \log p(y_t)\right) \\ + \log p(y_t) - \log p(v).\end{aligned} \tag{8}$$

Since v is given (fixed), $\log p(v)$ is a constant with respect to y_t and can be omitted, we simplify Equation (8) to:

$$\log \widehat{p}(y_t \mid v) \propto \lambda(\log p(y_t \mid v) - \log p(y_t)) + \log p(y_t) \tag{9}$$

Algorithm 1 Joint Training of LVLM with SIG

```
Require: $P$: Model; $\mathcal{X}, \mathcal{V}$: Training dataset
repeat
    $(\mathbf{x}, \mathbf{v}) \sim (\mathcal{X}, \mathcal{V})$    ▷ Sample multimodal input data
    $\mathbf{v} \leftarrow \epsilon$ with probability $\theta$    ▷ Randomly discard visual input, replace with learnable soft visual prompt $\epsilon$
    $\mathcal{L}_{\text{cross-entropy}} = -\mathbb{E}_{(\mathbf{x},\mathbf{v})} \sum_i \mathbf{y}_i \log P(\mathbf{x}, \mathbf{v})$
    Update $P$ and $\epsilon$
until converged
```

Equation (8) demonstrates that the influence of visual input v on response generation can be amplified by adjusting the scaling parameter λ, once given the conditional probability $p(y_t \mid v)$ of original inputs and the unconditional probability $p(y_t)$ without visual inputs. This formulation highlights a major challenge in enhancing visual guidance for LVLMs: accurately calculating the unconditional probability $p(y_t)$ of generating the token y_t in the absence of the visual input.

Previous approaches attempt to ascertain such probabilities probability by either providing the model with text-only input [66] or by injecting randomly generated noise to mask the image [22], thereby utilizing the model's output as the unconditional probability $p(y_t)$. Nonetheless, simply removing the visual inputs may disrupt input patterns(e.g., the input length and modalities), as visual tokens typically far surpass text tokens in quantity [6, 66]. Concurrently, adding random noise to distort images relies can introduce uncontrollable and unstable informational perturbations. The extra, unforeseen noise introduced by these inputs may lead the LVLMs to behave more like random probability generators, thereby complicating the approximation of $p(y_t)$.

Therefore, SIG first employs a learnable soft visual prompt ϵ to replace visual inputs, thereby forming a multimodal-null input for the model. The learnable soft visual prompt ϵ will be the jointly trained with the LVLM. As outlined in Algorithm 1, we replace visual input with ϵ with probability θ during training. The soft visual prompt ϵ serves dual purposes, acting both as a placeholder to maintain input patterns and as an indicator to make model prioritize text inputs. This dual functionality ensures a consistent input pattern for LVLMs in both training and inference, allowing the model to produce generate interpretable output and balancing visual and text inputs. After training, we can directly use the ϵ to query the model and extract the approximation of $p(y_t)$. Finally, during inference, we contrast output distributions from original and multimodal-null inputs based on Equation 9 to get the final output. Specifically, logits ℓ_g of generated tokens are recalculated by adjusting the logits ℓ_u of the multimodal-null inputs with the scaling parameter λ, based on logits ℓ_c of original inputs as follows:

$$\ell_g = \ell_u + (\ell_c - \ell_u) \times \lambda \tag{10}$$

Model	Model Size	Obj Hall		MMHall		LLaVABench	MM-VET
		Res ↓	Obj ↓	Score ↑	Hall ↓		
LRV [28]	7B	32.30	22.30	/	/	/	31.70
LLaVA-1.5 [30]	7B	46.71	25.08	2.19	59.00	64.40	31.10
VCD [22]	7B	47.40	25.24	2.12	59.00	65.30	30.90
VDD-None [66]	7B	46.71	25.19	2.22	56.00	66.00	31.70
Less-is-more [65]	7B	40.30	17.80	2.33	50.00	60.90	/
OPERA [16]	7B	45.10	22.30	2.15	54.20	60.30	/
HA-DPO [68]	7B	39.90	19.90	1.98	60.40	67.20	/
POVID [69]	7B	48.10	24.40	2.08	56.20	62.20	/
LLaVA1.5-7B-BPO [44]	7B	31.90	15.10	/	/	71.60	36.80
Ours	7B	**27.86**	**14.22**	**2.53**	**49.00**	**72.20**	35.20
Δ, compare to LLaVA-1.5	7B	40.36%	43.30%	15.53%	16.95%	12.11%	13.18%
LLaVA [29]	13B	63.00	29.50	/	/	70.80	26.40
Muffin [34]	13B	50.50	24.50	/	/	68.80	/
QWEN-VL [3]	10B	40.40	20.70	/	/	52.10	/
LLaVA-1.5 [30]	13B	47.06	23.33	2.54	50.00	72.50	36.10
VCD [22]	13B	46.37	23.10	2.60	49.00	73.60	36.90
VDD-None [66]	13B	44.64	22.23	2.38	55.00	73.00	36.10
LLaVA-RLHF [54]	13B	38.10	18.90	2.53	57.00	61.50	/
RLHF-V [63]	13B	12.20	7.50	2.45	51.00	51.40	/
LLaVA1.5-13B-BPO [44]	13B	27.30	12.90	/	/	74.40	41.40
Ours	13B	27.21	14.10	**2.65**	**48.00**	**84.30**	39.90
Δ, compare to LLaVA-1.5	13B	42.18%	39.56%	4.33%	4.00%	16.28%	10.53%

Table 1. Comparison of baselines across multiple benchmarks. The highest score marked in **bold** and the second highest underlined. The baselines are categorized into three groups: (1) typical LVLMs, indicated in **blue**, (2) training-free methods, indicated in **brown**, and (3) reinforcement learning-based methods, indicated in **gray**.

This refinement facilitates a more balanced and effective integration of visual inputs, thereby enhancing visual comprehension of LVLMs while addressing the language bias.

4. Experiments

4.1. Implementation Details

We strictly adhere to the settings of LLaVA-1.5 [30] for training the LVLM, which uses the same dataset and base model. The model is trained on 8 A100 GPUs, each with 40 GB of memory. The scaling parameter λ and replacement probability θ are determined based on the model's performance on the validation set. Our model shares almost the same structure and parameters as those of LLaVA-1.5, except for the proposed multimodal dual-attention mechanism and the learnable soft prompt used for soft-image guidance. More details is provided in the Appendix.

4.2. Evaluation Setup

We conduct experiments across three categories:

Visual Comprehension: **MMBench**[33] evaluates 20 fine-grained abilities of LVLMs, assessed with accuracy. **TextVQA** [53] employs VQA accuracy [1] as metric for question answering with text within images. Unlike LLaVA-1.5 [30], we do not provide the model with images OCR results. **MM-VET** [64] evaluates multimodal understanding with GPT-4 as free-form question-answering.

Open-ended Generation: LLaVA-Bench [29] uses GPT-4 to compare generated answers with reference answers.

Visual Hallucination: Object HalBench [48] detects object hallucinations by comparing model outputs with COCO labels [27]. We follow the same setup as [63], which adds diverse prompts with detailed image descriptions for stable evaluations. **MMHal-Bench** [54] evaluates hallucinations and response informativeness, with GPT-4 comparing model outputs to human responses and object labels.

4.3. Experimental Results

We evaluate the effectiveness of our proposed method across various benchmarks, as presented in Table 2. The baseline models are categorized into three groups: (1) LVLMs obtained after multimodal alignment training of foundational LLMs and the visual encoders, marked in **blue**; (2) training-free methods designed to mitigate hallucinations in LVLMs, marked in **brown**; and (3) reinforcement learning-based methods aimed at aligning LVLM outputs with human intentions, marked in **gray**.

Compared to these baseline models, our method demonstrates consistent performance improvements across all tasks and benchmarks. Notably, in comparison to LLaVA-

Method	Model Size	MMBench	TextVQA	LLaVABench	Obj Hall	
					Res ↓	Obj ↓
			Greedy Sampling			
LLaVA-1.5[30]	7B	64.61	46.05	64.40	46.71	25.08
VCD[22]	7B	64.69 (+0.08)	46.05 (+0.00)	65.30 (+0.90)	47.40 (+0.69)	25.24 (+0.16)
VDD-None[66]	7B	64.52 (-0.09)	44.47 (-1.58)	66.00 (+1.60)	46.71 (+0.00)	25.19 (+0.10)
w. SIG	7B	66.92 (+2.31)	46.77 (+0.72)	70.60 (+6.20)	30.36 (-16.35)	15.16 (-9.92)
LLaVA-1.5[30]	13B	67.74	48.66	72.50	47.06	23.33
VCD[22]	13B	68.38 (+0.64)	48.63 (-0.03)	73.60 (+1.10)	46.37 (-0.69)	23.10 (-0.23)
VDD-None[66]	13B	68.56 (+0.82)	47.31 (-1.35)	73.00 (+0.05)	44.64 (-2.42)	22.23 (-1.10)
w. SIG	13B	70.19 (+2.45)	48.74 (+0.07)	74.70 (+2.20)	28.27 (-18.79)	15.21 (-8.12)
			Nucleus Sampling			
LLaVA-1.5[30]	7B	56.96	35.41	63.00	56.66	29.75
VCD[22]	7B	60.91 (+3.95)	40.67 (+5.26)	65.30 (+2.30)	49.83 (-6.83)	27.44 (-2.31)
VDD-None [66]	7B	62.97 (+6.01)	42.62 (+7.21)	66.50 (+2.50)	57.34 (+0.86)	28.22 (-1.53)
w. SIG	7B	63.49 (+6.53)	39.40 (+3.99)	68.40 (+5.40)	29.14 (-27.52)	15.62 (-14.13)
LLaVA-1.5[30]	13B	62.11	38.92	68.10	50.52	25.74
VCD[22]	13B	65.38 (+3.27)	43.56 (+4.64)	70.70 (+2.60)	49.83 (-0.69)	24.23 (-1.51)
VDD-None[66]	13B	66.32 (+4.21)	45.99 (+7.07)	71.40 (+3.30)	47.90 (-2.62)	23.25 (-2.49)
w. SIG	13B	64.77 (+2.66)	40.31 (+1.39)	72.00 (+3.90)	30.55 (-19.97)	17.45 (-8.29)

Table 2. Comparison of SIG with training-free methods designed to mitigate hallucinations across different decoding strategies. Performance gap compared to the base model(LLaVA-1.5) are noted in parentheses. Red denotes improvements, while green indicates reductions.

1.5 [30], which shares the same training data and model structure with our method, our method achieves substantial double-digit percentage improvements, as indicated by the Δ values in Table 2. Specifically, we achieve a 40.36% reduction in response hallucination(Obj Hall) and a 12.11% improvement in LLaVA-Bench on the 7B model. On the 13B model, we also achieve a 42.18% reduction in response hallucination(Obj Hall) and an 16.28% improvement in LLaVA-Bench, demonstrating the robust scalability of our method with larger models. Our method also surpasses training-free methods such as VCD [22] and VDD [2], achieving nearly 20 points reduction on Object-Hallucination Benchmark. This demonstrates the superior capability of our method in mitigating visual hallucinations by addressing the language bias inherent in LVLMs. The underperformance of these methods further indicates that adding randomly generated noise on input images or simply remove the image during the inference injects the unexpected information that was not present during training, thereby diminishing the robustness of their methods across the benchmark. In comparison to reinforcement learning-based methods, which require extensive training resources, additional training data, and high-quality feedback data, our method remains efficient and cost-effective while delivering superior results. For example, compared to RLHF-V [63] and LLaVA1.5-BPO [44], our method achieves higher scores on most benchmarks without the need for reinforcement learning. Our 7B model attains a lower hallucination rate and higher visual comprehension scores than these methods, highlighting the efficiency of our proposed methods, including MDA and SIG.

4.4. Analysis Results

Effect of Soft-Image Guidance in Decoding Perspective To distinguish our methods from pervious works, we investigate effectiveness of SIG in different decoding strategies. Our analysis in Table 2 reveals previous training-free approaches, like VCD [22] and VDD-None [66], exhibit performance gains exclusively under Nucleus Sampling [14], whereas SIG shows consistently improvements in **Greedy Sampling** and **Nucleus Sampling** in all benchmarks.

VCD contrastes output distributions derived from original and distorted visual inputs. Similarly, VDD employs model's output based on text-only inputs as the contrastive source. However, adding randomly generated noise on input images or given text-only inputs during inference may introduce the unexpected information that was not present during training. This inconsistency can lead to suboptimal performance and reduced robustness of these methods. In benchmarks like MMBench and TextVQA where LVLMs generate limited words, this noise undermines the model's ability to generate reliable responses using **Greedy Sampling**. Greedy Sampling deterministically selects the most probable token at each step, offering minimal opportunity to counteract the introduced noise, specially in short outputs. Consequently, Nucleus Sampling select tokens by probability, thereby mitigating the inherent instability introduced by these methods. Nevertheless, the Nucleus Sampling itself may hinder model performance on the benchmarks with constrained settings, resulting in a noticeable decline in per-

Model	LLaVABench			
	Complex	Conv	Detail	All
LLaVA-1.5	75.50	54.10	56.60	64.40
w. FastV [6]	79.80	54.10	46.70	63.90
Δ	+4.30	+0.00	-9.90	-0.50
MDA	83.20	59.70	59.20	70.30
w. FastV [6]	10.70	10.20	10.40	10.50
Δ	**-72.50**	**-49.50**	**-48.80**	**-59.80**

Table 3. Performance comparison on LLavaBench between standard LLaVA-1.5 and those with MDA, with and without FastV [6].

Sampling	Model	MMBench	TextVQA	LLavaBench	MM-VET
Greedy	LLaVA-1.5	64.61	46.05	64.40	31.10
	Ours	65.98	**46.94**	**72.20**	**33.50**
	-w/o. MDA	**66.92**	46.77	70.60	32.00
	-w/o. SIG	65.46	46.03	70.30	32.80
Nucleus	LLaVA-1.5	56.96	35.41	63.00	29.80
	Ours	**64.69**	**42.05**	**72.20**	**35.20**
	-w/o. MDA	63.49	39.40	68.40	33.30
	-w/o. SIG	56.62	36.40	67.80	30.50

Table 4. Ablation study on under different decoding strategy across multiple benchmarks on 7B model.

formance compared to Greedy Sampling.

In contrast, SIG employs a learnable soft visual prompt to replace visual inputs, which serves as a placeholder that maintains input patterns while suggesting the model to prioritize the text input. SIG keeps consistency for the model inputs during training and inference and improves the overall performance under both decoding strategies.

How does LVLMs treat visual inputs with Multimodal Dual-Attention Mechanism? To delve into the effectiveness of the MDA in mitigating language bias caused by disparities in training scales, we examine the behavior of LVLMs regarding visual inputs across all layers. To assess whether MDA addresses this issue, we adopt the pruning method proposed by Chen et al. [6] on LLaVA-1.5 with MDA. Specifically, we prune half of visual tokens in deeper layers and subsequently examining the impact on model performance across LLaVABench. Previous studies Chen et al. [6] on LVLMs, such as LLaVA-1.5, have demonstrated that pruning visual tokens in deeper layers does not greatly affect the model performance, indicating that deeper layers are not utilizing these tokens effectively. Conversely, our experiments in Table 3 reveal that employing FastV in the model with MDA leads to a substantial performance drop. This finding indicates that the model is indeed leveraging visual inputs across all layers, not just in the shallow ones. The significant performance reduction confirms that MDA effectively forces the LVLM to utilize visual inputs more thoroughly. By ensuring comprehensive attention to visual inputs across the model's layers, MDA facilitates the LVLM in fully exploiting its visual comprehension capabilities. The 7.7-points improvement for complex tasks on LLaVABench, as shown in Table 3, further corroborates this conclusion, as complex tasks generally require the deeper layers for precise understanding [4, 18].

Ablation Study To better understand the contributions of each component in our method, we conduct an ablation study across multiple benchmarks. We report results under different decoding strategies on the 7B model in Table 4. Removing MDA (" w/o MDA ") causes a significant drop in performance, particularly on LLavaBench and MM-VET. This suggests that MDA is crucial for enabling the model to effectively integrate visual information across the model. Excluding the SIG (" w/o SIG ") also leads to a notable performance decrease across all benchmarks. Interestingly, SIG-alone (" w/o MDA ") yields better performance on MMBench under greedy decoding, likely due to the evaluation of this benchmark in multiple-choice QA. Since SIG modifies output distribution, even slight changes can improve the likelihood of reaching at correct answers by chance, particularly when the model struggles to comprehend the inputs and assigns random probabilities to the answer options for decoding. In contrast, with the combination of MDA, the model tend to focus on visual inputs, potentially stabilizes the model‘s output distribution. It makes the slight variations from SIG less influential, thereby reducing the likelihood of obtaining correct answers and leading in suboptimal performance in this benchmark. Nevertheless, both components individually contribute to substantial improvements over the baseline LLaVA-1.5 model. Even when one component is removed, the model still outperforms the baseline in all benchmarks.

5. Conclusion

This paper tackles the language bias in LVLMs, which often leads to the neglect of visual inputs and the generation of hallucinatory responses. We identify two primary sources of this bias: the gap in training scales between the pretraining and multimodal alignment stages of LVLMs, and the learned inference bias due to the short-term dependency in text data. To address these challenges, we introduced the Multimodal Dual-Attention Mechanism (MDA) and Soft-Image Guidance (SIG). MDA enhances the integration of visual inputs across all layers. SIG proposes a novel decoding strategy to mitigate over-reliance on adjacent text tokens, using a learnable soft visual prompt. Our work highlights the importance of addressing language biases from both training and inference perspectives, paving the way for more accurate and reliable LVLMs.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. 6

[2] Ibrahim Alabdulmohsin, Vinh Q. Tran, and Mostafa Dehghani. Fractal patterns may illuminate the success of next-token prediction, 2024. 1, 7

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1, 6, 12, 13

[4] Amit Ben-Artzy and Roy Schwartz. Attend first, consolidate later: On the importance of attention in different llm layers, 2024. 8

[5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. 1, 3, 12

[6] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models, 2024. 1, 2, 3, 4, 5, 8, 12

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023. 1

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 12

[9] Michał Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. Frustratingly short attention spans in neural language modeling, 2017. 1

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 3

[11] Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, et al. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *arXiv preprint arXiv:2410.16261*, 2024. 3

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 12

[13] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 1

[14] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. 7

[15] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. 3

[16] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation, 2024. 6, 14

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 12

[18] Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. Exploring concept depth: How large language models acquire knowledge at different layers?, 2024. 8

[19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 12

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 12

[21] Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. A survey of hallucination in large visual language models, 2024. 1, 3

[22] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. 2, 3, 5, 6, 7, 14

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 3

[24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023. 1

[25] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 3

[26] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models, 2024. 1

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6, 13

[28] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2024. 3, 6, 13, 14

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 3, 6, 12, 13, 14

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 3, 4, 6, 7, 12, 13, 14, 15

[31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1

[32] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024. 3

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 6, 13, 15

[34] Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. Muffin: Curating multi-faceted instructions for improving instruction-following, 2024. 6, 13, 14

[35] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 12

[36] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt, 2023. 1

[37] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 12

[38] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024. 1, 3

[39] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 12

[40] OpenAI. Introducing chatgpt. 2022. 1

[41] OpenAI. Gpt-4v(ision) system card. 2023. 1, 3

[42] OpenAI. Gpt-4 technical report, 2023. 1, 3

[43] OpenAI. hello-gpt-4o, 2024. 1, 3

[44] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization, 2024. 1, 2, 3, 6, 7, 14

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763, 2021. 12

[46] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. 14

[47] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 12

[48] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. 3, 6, 13

[49] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 12

[50] ShareGPT. `https://sharegpt.com/`, 2023. 12

[51] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 12

[52] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering, 2024. 1

[53] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. 6, 13

[54] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. 3, 6, 13, 14

[55] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*, 2023. 1, 3

[56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, 2023. 1

[57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288, 2023. 1

[58] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[59] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding, 2024. 3

[60] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024. 1

[61] Ruiqing Yan, Linghan Zheng, Xingbo Du, Han Zou, Yufeng Guo, and Jianfei Yang. Recurformer: Not all transformer heads need self-attention, 2024. 1

[62] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[63] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024. 3, 6, 7, 13, 14

[64] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. 6, 13, 14

[65] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective, 2024. 6, 14

[66] Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing multimodal large language models, 2024. 1, 2, 3, 4, 5, 6, 7, 14

[67] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning, 2024. 1, 3, 12

[68] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2024. 6, 14

[69] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning, 2024. 6, 14

[70] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2024. 3

[71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 3

[72] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding, 2024. 3

Appendix

This appendix is organized as follows.

- In Section A, we discuss the limitations for LACING.
- In Section B, we show implementation details for LACING. Specifically, training details(Sec. B.1), datasets(Sec. B.2) and hyperparameters(Sec. B.3).
- In Section C, we present the details of our experiments and evaluation. Specifically, dataset and metric(Sec. C.1), baselines(Sec. C.2) and GPT-4 Version(Sec. C.3)
- In Section D, we detail the experiments and provide an in-depth discussion on the impact of hyperparameters, specifically the replace probability θ(Sec. D.1) and the scaling parameter λ(Sec. D.2).
- In Section E, we present a human evaluation of LACING versus LLaVA-1.5 across LLaVABench.
- In Section F, we present more qualitative results.
- In Sectionn G, we visualized the attention distribution across different layers in LLaVA-1.5 and LACING.

A. Limitation

Despite the promising results demonstrated by LACING in addressing the language bias of LVLMs, several limitations must be acknowledged. First, our method is mainly conducted on the LLaVA-Family model [29, 30]. The effectiveness of our approach on other LVLMs with different architectures remains largely unexplored. Nevertheless, language bias is commonly observed across various LVLMs [5, 6, 67]. Both the Q-former-based and Resampler-based methods [3] are susceptible to language bias, exhibiting several phenomena similar to those seen in the LLaVA model [6]. Therefore, inspired by this common motivation, we anticipate the implementation and effectiveness of LACING on diverse LVLMs. Additionally, due to resource constraints, we are unable to acquire LVLMs that achieve a similar scale of training between the LLM pretraining stages and the LVLM alignment stage to accurately validate the source of language bias. Finally, while LACING has significantly reduced hallucinations in LVLMs and enhanced visual comprehension capabilities, there remains a possibility for it to produce hallucinations or disseminate misinformation. Therefore, it still should be employed with caution in critical applications. Consequently, future research could involve broadening our approach to include a wider spectrum of LVLMs with different architectures and training them using a comparable training scale to observe the manifestations of language bias.

B. Training Details

To make fair compression, we adopt the same training settings as LLaVA-1.5 [30], maintaining consistency in hyperparameters, training dataset, data preprocessing, and model architecture. The only differences lies in the introduction of the multimodal dual-attention mechanism and the learnable soft visual prompt for soft-image guidance.

Dataset	Data Size
LLaVA [29]	158K
ShareGPT [50]	40K
VQAv2 [12]	83K
GQA [17]	72K
OKVQA [37]	9K
OCRVQA [39]	80K
A-OKVQA [49]	66K
TextCaps [51]	22K
RefCOCO [19, 35]	48K
VG [20]	86K
Total	665K

Table 5. Instruction-following Data Mixture Used for Finetuning [30].

B.1. Training

Following the setting of LLaVA-1.5 [30], we employ CLIP-ViT-L-14-336 [45] as the visual encoder, paired with a two-layer MLP adapter to project visual embeddings from the encoder to the LLM backbone. Vicuna-1.5 [8] serves as the LLM backbone. All of the experiments are conducted on the 8 × A100 GPUs, each with 40 GB of memory. We employ the Deepspeed Zero2 [47] and Deepspeed Zero3 [47] for training the 7B and 13B model, respectively.

In addition to these standard components of LLaVA-1.5, our method includes two significant modifications to the model architecture. Firstly, we adopt the multimodal dual-attention mechanism proposed in this paper, replacing the vanilla self-attention in the LLM. This modification slightly increases the computational cost due to the dual-attention calculation. We further incorporate a learnable soft visual prompt for soft-image guidance. We maintain a learnable embedding with dimensions $[l_{\text{visual}}, h_{\text{LLM}}]$, where l_{visual} is the visual embedding length and h_{LLM} is the LLM hidden state size. In our practice, the learnable soft visual prompt has a size of $[576, 4096]$ for a 7B model and $[576, 5120]$ for a 13B model, which correspondingly adds 2.36M and 2.95M parameters to the 7B and 13B models. Compared to the billion-level parameters of these LVLMs, the additional parameters account for only 0.03% and 0.02%, respectively, which are minimal and negligible. Therefore, compared to LLaVA-1.5, our method does not require additional training resources or computational costs, thereby demonstrating the efficiency of our approach. Practically speaking, the time cost of our method is approximately identical to that of LLaVA-1.5 under the same setting.

B.2. Data

We strictly follows the data setting of LLaVA-1.5 for both pretraining and finetuning. Specifically, the LLaVA-

Hyperparameter	Pretrain	Finetune
batch size	256	128
lr	1e-3	2e-5
lr schedule	cosine decay	cosine decay
lr warmup ratio	0.03	0.03
weight decay	0	0
optimizer	AdamW	AdamW
DeepSpeed stage	2	3
replace prob. θ	10%	10%

Table 6. **Hyperparameters** of LACING, which are the same as the original LLaVA-1.5 [30], except that we set the replace probability θ for training with soft-image guidance.

558K [29] for pertraining and a mixture of instruction-following data for finetuning shown in Table 5.

B.3. Hyperparameters

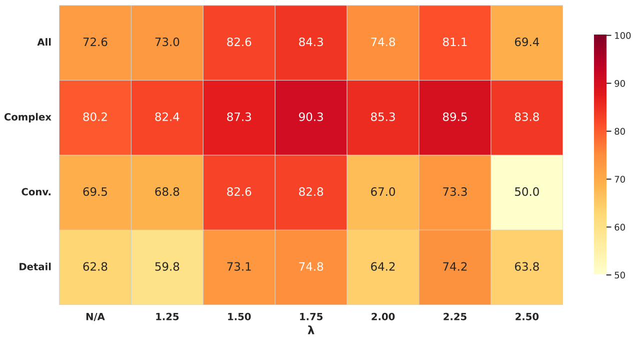


Figure 5. Model performance on LLaVABench across various scaling parameter λ.

We utilize the identical set of hyperparameters as the original LLaVA-1.5 [30], with the exception of specifying the replacement probability θ for training with soft-image guidance. Detailed training hyperparameters for both stages are provided in Table 6. During the inference, we use the hyperparameter λ to control the guidance of the visual inputs on the response generation. As illustrated in Figure 5, we report the performance of the 13B model on LLaVABench across various the scaling parameter λ, thereby demonstrating the impact of different λ scales on model performance. The optimal performance of our method under various λ values is reported in the experiments.

C. Detailed Experimental Settings

C.1. Dataset and Metric

MMBench [33] provides a progressive evaluation framework, advancing from perception to reasoning, and covers 20 fine-grained abilities. It is assessed through multiple-choice question answering, using accuracy as the metric.

MMBench [33] provides a progressive evaluation framework, advancing from perception to reasoning, and covers 20 fine-grained abilities. It is assessed through multiple-choice question answering, using accuracy as the metric.

TextVQA [53] is designed for visual question answering involving text within images. It employs VQA accuracy as the evaluation metric. Unlike LLaVA-1.5[30], which includes OCR results of the images in the question, our approach provides the model solely with the image and the question. This setup aims to assess the model's visual comprehension abilities without supplementary textual data.

MM-VET [64] evaluates multimodal understanding across six core vision-language capabilities over 128 tasks. The evaluation is conducted using GPT-4 to assess model performance in a free-form question-answering format. MM-Vet defines 16 integrations derived from combinations of these core capabilities, providing a structured assessment of models' abilities to handle complex multimodal tasks.

LLaVABench [29] is utilized for evaluating open-ended generation capabilities. This benchmark consists of 60 tasks focused on LLaVA's visual instruction-following and question-answering abilities in natural environments. It employs GPT-4 as the evaluator to compare the model's generated answers with reference answers, ensuring a comprehensive assessment of the model's generative performance.

Object HalBench [48] detects object hallucinations by comparing model outputs with COCO image labels [27]. Yu et al. [63] further augment this benchmark by adding eight diverse prompts with detailed image descriptions for stable evaluations. We follow the same evaluation setup and use GPT-4 as the evaluator. We report the two metrics in this benchmark: The response-level hallucination rate and the object-level hallucination rate.

MMHal-Bench [54] evaluates hallucinations and response informativeness. It employs GPT-4 to compare model output with human response and several object labels to get the final scores.

C.2. Baselines

General LVLMs that have undergone multimodal alignment training. Specifically, we utilize LLaVA [29], Qwen VL [3], LLaVA-1.5 [30], Muffin [34], and LRV [28] as representative baselines. These LVLMs are predominantly trained with multimodal data for alignment [3, 29, 34]

Model	MMBench	TextVQA
Greedy Sampling		
LLaVA-1.5 [30]	64.61	46.05
-w. Two epoch	65.63	45.83
w. SIG	66.92	46.77
-w. Two epoch	66.58	47.15
Nucleus Sampling		
LLaVA-1.5 [30]	56.96	35.41
-w. Two epoch	60.82	36.70
w. SIG	63.49	39.40
-w. Two epoch	62.97	41.27

Table 7. Performance comparison of models undergoes training for one or two epochs across MMBench and TextVQA.

and fine-tuned using high-quality instruction data [28, 30], thereby achieving exceptional performance in various multimodal tasks. For example, LRV [28] employs supervised fine-tuning on an expertly crafted visual preference dataset to mitigate hallucinations in LVLMs. Typically, these models integrate a pre-trained visual encoder with a large language model through an alignment module.

Training-free methods designed to mitigate hallucination of LVLMs. VCD [22] contrast model outputs generated from original inputs and distorted visual input to reduce over-reliance on statistical bias and unimodal priors. Similiarly, VDD [66] contrast model outputs from original inputs and inputs without visual inputs to reduce the influence of textual LLMs. OPERA [16] introduces a penalty term on the model logits during the beam-search decoding to mitigate the over-trust toward a few summary tokens. Less-is-more [65] proposes a selective end-of-sentence (EOS) special token supervision loss coupled with a data filtering strategy to improve the model's capacity for timely termination of generation, thereby mitigating hallucinations.

Reinforcement Learning-based method aimed at aligning LVLM outputs with human intentions to mitigate hallucination of LVLMs. Specifically, POVID [69] addresses hallucinations in VLLMs using AI-generated feedback. It first prompts GPT-4V to add hallucinations to correct answers and use distorts images to invoke the VLLM's inherent hallucination tendencies. The model is then trained with this generated data using direct preference optimization approaches [46] to mitigate hallucinations. HA-DPO [68] propose a pipeline for constructing positive and negative sample pairs and adopt the direct preference optimization [46] using the constructed dataset to reduces hallucination. RLHF-V [63] employs the Muffin [34] as the LLM backbone and collects 1.4k fine-grained correctional human feedback. The model is trained using this dataset through the proposed dense direct preference optimization method to reduce hallucination. LLaVA-BPO [44] proposes a pipeline to gather preference datasets and conduct preference learning to mitigate this type of hallucination.

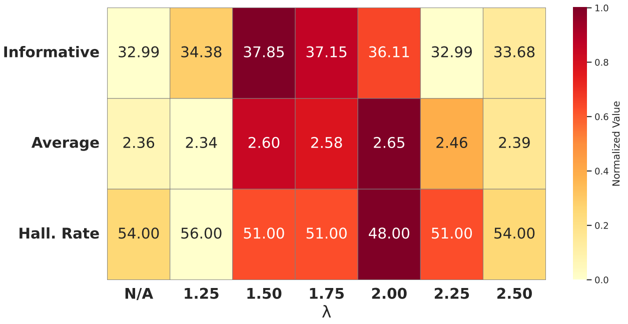


Figure 6. Model performance on MMHal-Bench across various scaling parameter λ.

C.3. GPT-4 Version

For all evaluations conducted using the GPT-4(evaluation for Object HalBench, MMHal-Bench, LLaVABench, and MM-VET), we utilized the `GPT-4 API` in October 2024. It ensures consistency with prior research [29, 54, 63, 64]. According to the documentation provided by OpenAI[1], `GPT-4 API` currently points to `GPT-4-0613 API`.

D. Parameter Study

D.1. Influence of the Replace Probability θ

In the soft-image guidance we proposed, we intermittently replace the visual input with a learnable soft visual prompt at a predetermined probability rate to give the model an input without visual input during training. This introduces segments of training data that remain unseen by the LVLMs during training. Consequently, we make the model that undergoes training for two epochs as a baseline to ensure comprehensive exposure to all samples in the training dataset. Subsequently, we evaluate the model after one and two epochs of training on the same benchmarks to determine the impact of visual input replacement. The results presented in Table 7 indicate that neither the number of training epochs nor the visual input replacement significantly impacts model performance, as it remains consistent across various settings and does not exhibit a clear trend of performance variation related to different training settings. To further establish the appropriate value of the replace probability θ, we present an experiment in Table 8 to identify the optimal value for this parameter.

---

[1] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

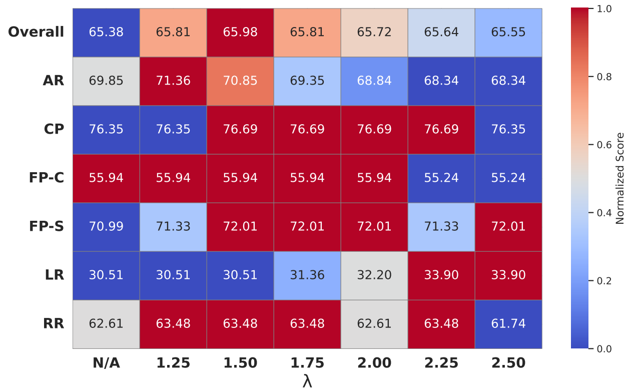


Figure 7. Model performance on MMBench across various scaling parameter λ.

θ	5%	10%	15%	20%
MMBench [33]	66.32	**66.92**	66.75	65.64
LLaVABench [33]	67.00	**70.60**	67.80	66.90

Table 8. Performance of SIG on MMBench and LLavaBench across different replace probability θ

D.2. Impact of the Scaling Parameter λ

Another essential hyperparameter is the scaling parameter λ, which is employed in soft-image guidance to regulate the guidance of the visual inputs towards the response generateion. Therefore, To assess the effect of varying λ values comprehensively, we examine our method's performance on MMBench, LLaVABench and Hall-Bench with different λ values, which can be divided into two distinct scenarios: multi-choice generation and open-end generation. The experimental results, illustrated in Figure 7, Figure 5, and Figure 6, suggest that an optimal value for the scaling parameter λ lies between 1.5 and 2.0. This range provides suitable visual guidance without impairing the text generation capabilities of LVLMs.

E. Human Evaluation on LLaVABench

To better illustrate the efficacy of our method, a further human evaluation has been undertaken to compare the model performance of LACING versus LLaVA-1.5 [30]. Specifically, we evaluate the model perofrmance on LLaVABench, which consists of 60 instances. We invited three human participants (all of them are Ph.D. students or Master students) to compare the responses generated by the models. For each comparison, three options were provided (Win, Tie, and Lose), with the final results determined by the majority vote of the participants. Figure 8 showcases the effectiveness of our method.

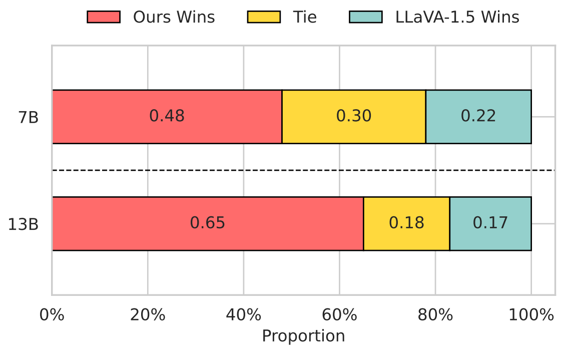


Figure 8. Human evaluation on LLaVABench.

During the human evaluation, the participants adhere the following principles to make the decision:

Principles of Human Evaluation for LLaVABench

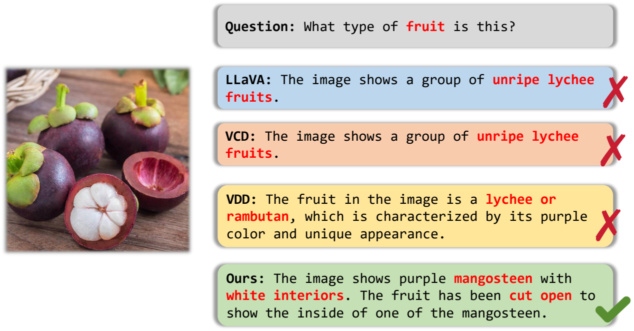
You are asked to evaluate the responses generated by different models. Your evaluation should adhere to the following principles:
1.**Correctness**: Assess whether responses address the key points outlined in the reference answer and image. For reference answers with multiple key points, evaluate how many of these the response accurately addresses and score accordingly. Additionally, ensure that the response provides the necessary information for the user.
2.**Faithfulness**: Examine any additional information in the answer to verify its accuracy and relevance to the question and image. If this information is incorrect or not relevant to the question and image, points should be deducted.
4.**Coherence**: Evaluate the fluency and coherence of the responses. Also, consider deducting points for overly verbose responses or those that are excessively generalized.

Finally, please make a decision among 3 opinions, including Win, Tie, and Loss.

If the majority voting of three participants not yield a decisive outcome, we will engage in further discussions among the involved participants and subsequently conduct another vote to determine the final result.

F. Case Studies

To deliver a thorough evaluation of the effectiveness of our methods in mitigating visual hallucinations and enhancing the visual comprehension of LVLMS, we present a case study in this section. We compare the open-ended generation results of our methods against several baseline models



Question: What type of fruit is this?

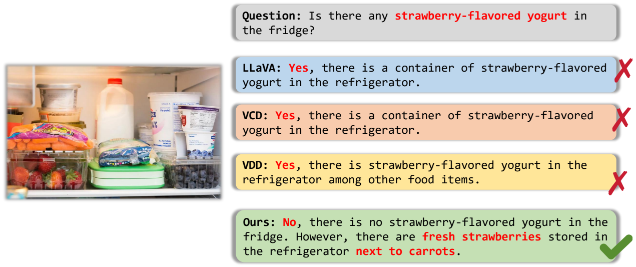
LLaVA: The image shows a group of unripe lychee fruits. ✗

VCD: The image shows a group of unripe lychee fruits. ✗

VDD: The fruit in the image is a lychee or rambutan, which is characterized by its purple color and unique appearance. ✗

Ours: The image shows purple mangosteen with white interiors. The fruit has been cut open to show the inside of one of the mangosteen. ✓

Figure 9. Comparison of LACING with other baselines on a sample from LLaVABench. Hallucinated responses and our corrections are highlighted in **red**.



Question: Is there any strawberry-flavored yogurt in the fridge?

LLaVA: Yes, there is a container of strawberry-flavored yogurt in the refrigerator. ✗

VCD: Yes, there is a container of strawberry-flavored yogurt in the refrigerator. ✗

VDD: Yes, there is strawberry-flavored yogurt in the refrigerator among other food items. ✗

Ours: No, there is no strawberry-flavored yogurt in the fridge. However, there are fresh strawberries stored in the refrigerator next to carrots. ✓

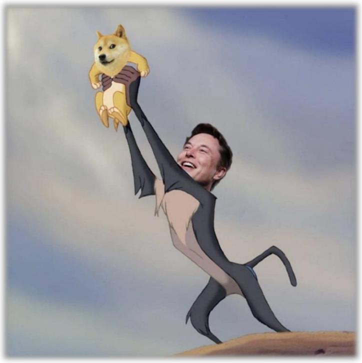
Figure 10. Comparison of LACING with other baselines on a sample from LLaVABench. LACING demonstrates a reduction in object hallucination and an enhancement in fine-grained visual comprehension, such as the identification of fresh strawberries in the refrigerator.

utilizing samples from LLaVABench. The evaluations of the case studies on the 13B model are illustrated in Figure 9, Figure 10, Figure 11, and Figure 12.

G. Comparison of Attention Allocation with Standard LVLMs

To provide a clear illustration of our proposed multimodal dual-attention mechanism and distinguish the attention allocation towards visual tokens and text tokens between our model and the standard LVLM, we visualized the attention distribution across different layers in LVLMs in Figure 13 and Figure 14, comparing the standard LVLM (LLaVA-1.5) with MDA. We get the attention allocation of LVLMs shown in this paper, such as Figure 13, and Figure 14, using the data from LLaVABench. We select 30 samples from the LLaVABench to query the LVLMs and examine the attention allocation during generation. We report the avg scores in Figure 2 and Figure 4.

Question: Which **iconic movie scene** is being parodied in the meme?



LLaVA: The meme is parodying the iconic scene from the movie "The Lion King" where Simba is lifted up by Rafiki, who holds him above the Pride Rock. In the meme, **a man** is holding a small dog, and the dog is positioned in a way that it appears to be lifted up by the man, mimicking the scene from "The Lion King." This humorous and creative adaptation of the original scene adds a playful twist to the image. ✗

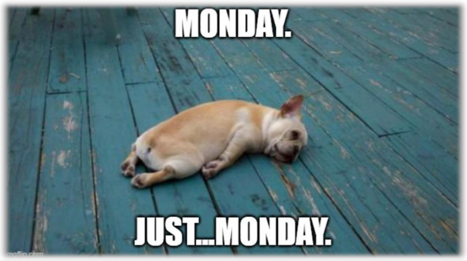
VCD: The meme is parodying the iconic scene from the movie "The Lion King" where Simba is lifted up by Rafiki, who holds him above the Pride Rock. In the meme, **a man** is holding a small dog, and the dog is positioned in a way that it appears to be lifted up by the man, mimicking the scene from "The Lion King." This humorous and creative adaptation of the original scene adds a playful twist to the image. ✗

VDD: The meme parodies the iconic scene from the movie "The Lion King" where **Simba holds a baby animal**, likely a lion cub, in his hands. In this particular meme, a man, possibly Elon Musk, is **portrayed as the character Simba**, holding a small dog in his arms. This creative and humorous adaptation of the original scene adds a modern twist by incorporating a well-known person, Elon Musk, into the parody. ✗

Ours: The meme in the image is a parody of a scene from the Lion King, where Rafiki holds up a baby lion while Simba adoringly looks up at him. **Elon Musk is portrayed as Rafiki**, **smiling** as he **looks up at the puppy** he's holding. The meme has skillfully combined elements from the Lion King and a photograph of Elon Musk to create an amusing and entertaining scene. ✓

Figure 11. Comparison of LACING with other baselines on a sample from LLaVABench. LACING effectively recognizes the individual in the meme and provides a thorough description, guiding to accurate responses to the question.

Question: Please explain the **meme** in detail.



LLaVA: The image features a small dog lying on a wooden floor, possibly a deck, with a caption above it. The caption reads **"Monday. Just Monday. . ."** The dog appears to be **relaxed and enjoying** its time on the wooden surface. The scene captures a moment of **leisure and contentment for the dog**, as it spends its day indoors. ✗

VCD: The image features a small brown dog lying on a wooden floor, appearing to be sleeping or resting. The dog is positioned in the center of the scene, occupying a significant portion of the image. The floor appears to be made of wood, giving the scene a cozy and warm atmosphere. The image is accompanied by a caption that reads, **"Monday. Just... Monday."** This caption adds a humorous touch to the scene, as it suggests that the dog is expressing its feelings about the start of the week. The combination of the dog's relaxed posture and the caption creates a **lighthearted and amusing meme**. ✗

VDD: The image features a small brown dog lying on a hardwood floor on its side, taking up a significant portion of the floor. The caption above the dog reads **"Monday. Just... Monday!"**. In the background, there **are chairs placed at various positions**, adding to the overall setting of the scene. The combination of the dog's playful position and the caption's clever wordplay creates a **light-hearted and relatable meme**. ✗

Ours: The image is a **Monday motivation meme** featuring a small tan dog lying on a hardwood floor. The caption reads, **"Monday. Just... Monday."** The meme expresses the **sentiment of feeling down or unenthusiastic** about the start of the work week, as mondays are often associated with work, school, and other responsibilities. The dog in the image symbolizes a sense of relaxation and leisure, **contrasting with the more stressful start of the week**. ✓

Figure 12. Comparison of LACING with other baselines on a sample from LLaVABench. LACING is the only model capable of successfully articulating the idea that the meme is trying to convey, by contrasting image information and questions.

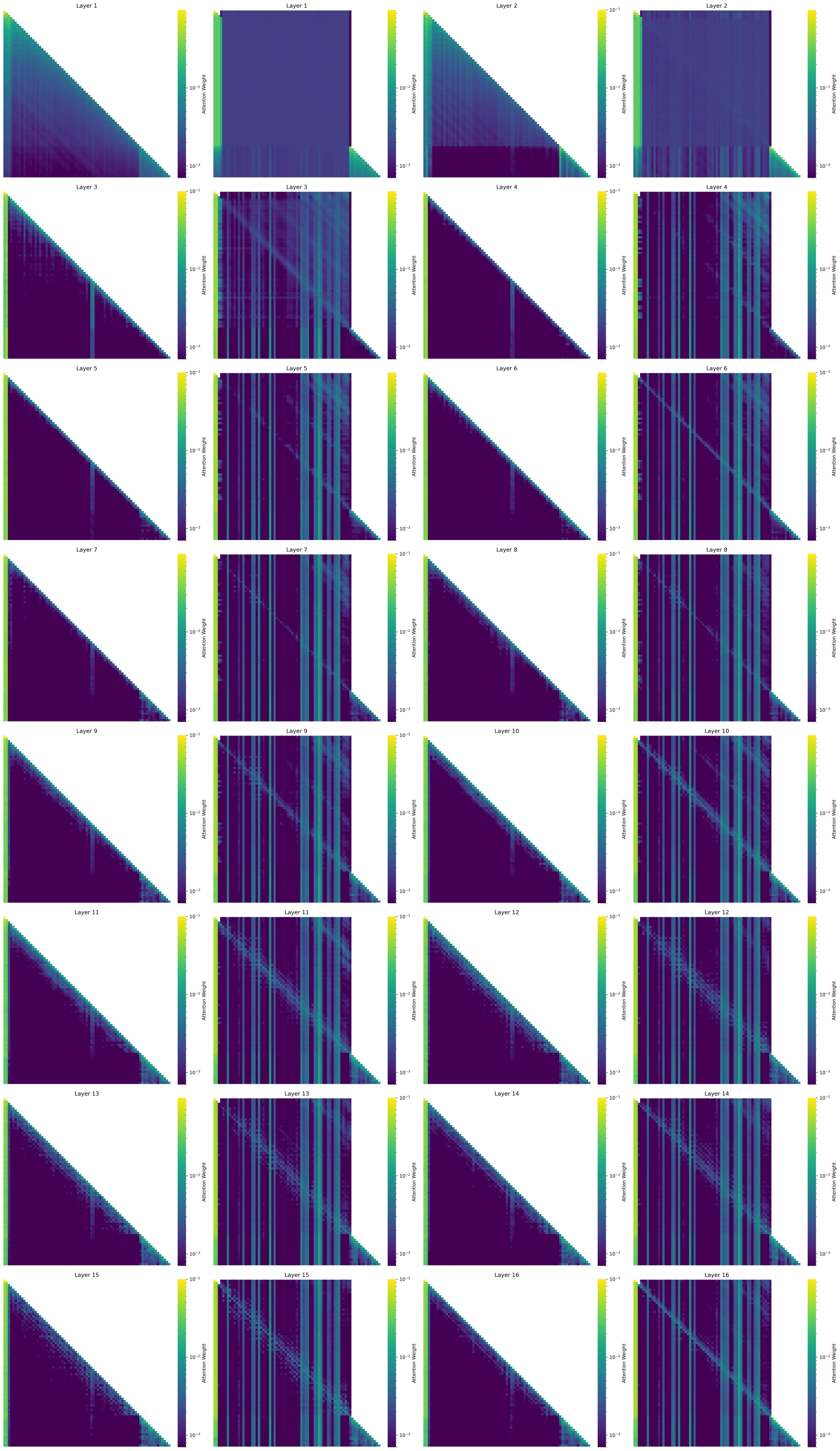


Figure 13. Comparison of Attention Maps across the 1st to 16th Layer in LLaVA and LACING.

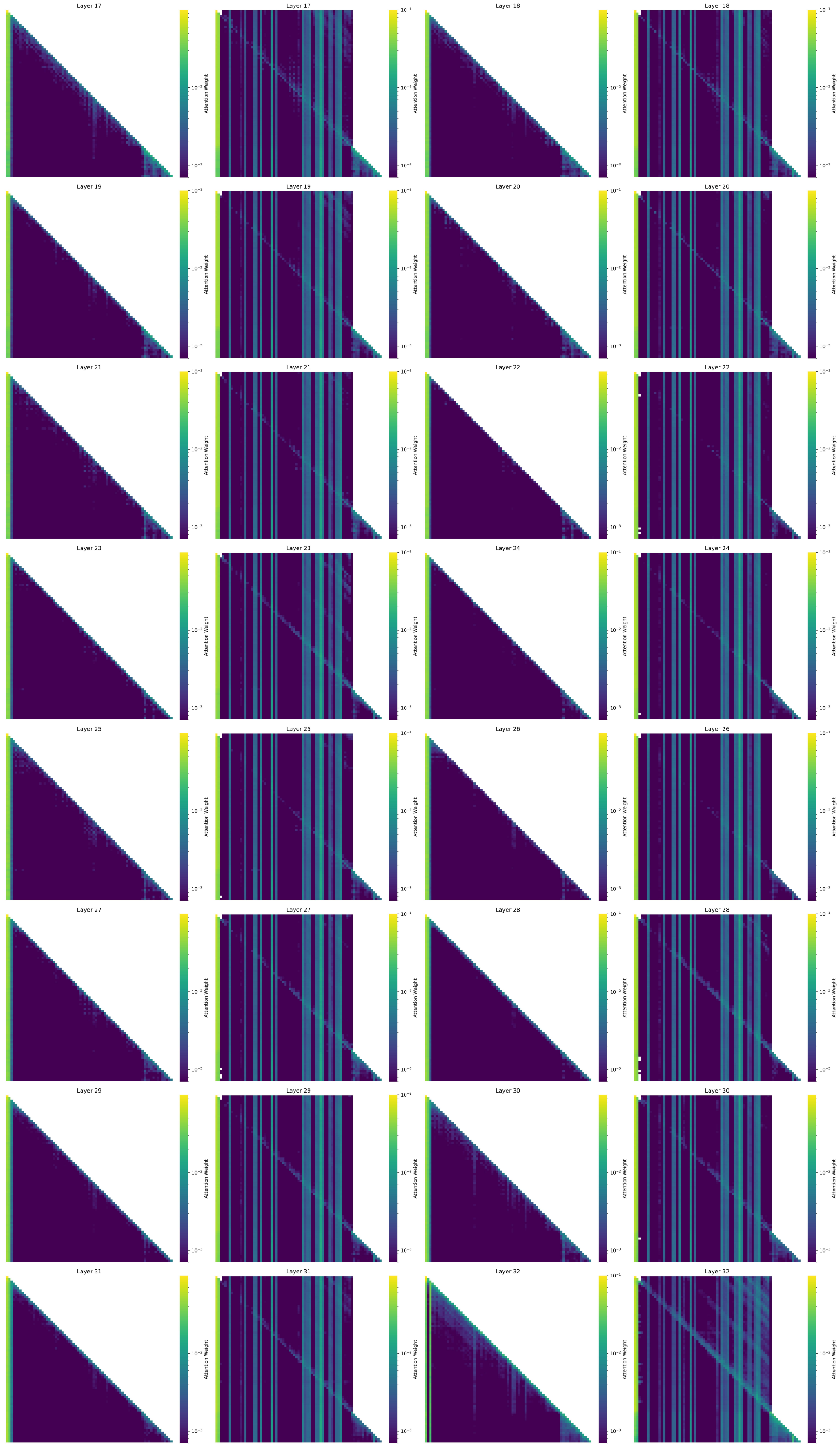


Figure 14. Comparison of Attention Maps across the 17th to 32nd Layer in LLaVA and LACING.