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

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) have achieved impressive results in vision-language tasks. However, LVLMs suffer from hallucinations caused by language bias, which neglects images while over-relying on text. We identify two reasons for the bias: 1). Different training scales between the LLM pretraining and LVLM alignment stage. 2). The learned inference bias due to short-term dependency of text data. Therefore, we propose LACING, designed to address such bias with MuLtimodal DuAl-attention MeChanIsm (MDA) aNd Soft-Image Guidance (SIG). Specifically, MDA adopts a **parallel dual-attention mechanism** that constructs separate attention for visual and text inputs to enhance integration of visual inputs across model. SIG uses a **learnable soft visual prompt** during training and inference to replace visual inputs, designed to compel LVLMs to prioritize text inputs during inference. Experiments across different model architectures and scales demonstrate that LACING effectively debiases LVLMs from their language bias, enhancing visual comprehension and reducing hallucinations without additional resources.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2023; Dubey et al., 2024) represent a significant milestone in natural language processing (Yang et al., 2024; OpenAI, 2022). By incorporating visual encoders into LLMs (Liu et al., 2023; Bai et al., 2023), the development of Large Vision-Language Models (LVLMs) (OpenAI, 2024; Team, 2023) has been accelerated, enabling them to handle both visual and text inputs. This facilitates various applications using LVLMs such as autonomous driving (Xu et al., 2024) and medical assistants (Li et al., 2023b).

State-of-the-art LVLMs, despite their advanced capabilities in handling both modalities, often produce erroneous or irrelevant responses to input images (Chen et al., 2024b; Lan et al., 2024). he main reason behind such hallucinations is referred to as language bias (Zhao et al., 2024a), i.e., models sometimes "ignore" visual inputs and generate text responses solely based on text inputs. However, prior studies have not comprehensively explored the origins of such bias. We suggest that this bias may emerges for the following two reasons:

1. Different training scales between pretraining and multimodal alignment stage: The LLM backbone in LVLMs is pre-trained on on extensive text corpus, while the multimodal alignment stage of LVLMs involves significantly fewer samples and shorter training duration. For instance, Llama3 (Dubey et al., 2024) is pre-trained with 15T tokens, whereas the multimodal alignment training for LLaVA-Series (Liu et al., 2023, 2024c,d) employs only about 558k–1.3M examples. This scale discrepancy causes the pretraining distribution to dominate the generation process in LVLMs (Pi et al., 2024), resulting in insufficient utilization of visual inputs. As shown in Figure 2, LVLMs allocate minimal attention to visual tokens in over 90% layers (Chen et al., 2024a). Conversely, as discussed in § E, models such as Chameleon (Team, 2024), pretrained with balanced scales of textual and visual tokens, exhibit significantly reduced bias, further supporting this hypothesis.

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 (Alabdulmohsin et al., 2024; Daniluk et al., 2017; Yan et al., 2024), i.e., the short-term dependency of text data. LLMs pre-trained on large-scale text corpora are more easily capturing and memorizing such short-term dependency (Yuan et al., 2025), which typically assign higher attention weights to adjacent tokens. However, this learned pattern may be problematic in multi-modal contexts. In current LVLMs, visual features are typically concatenated with text inputs to form input

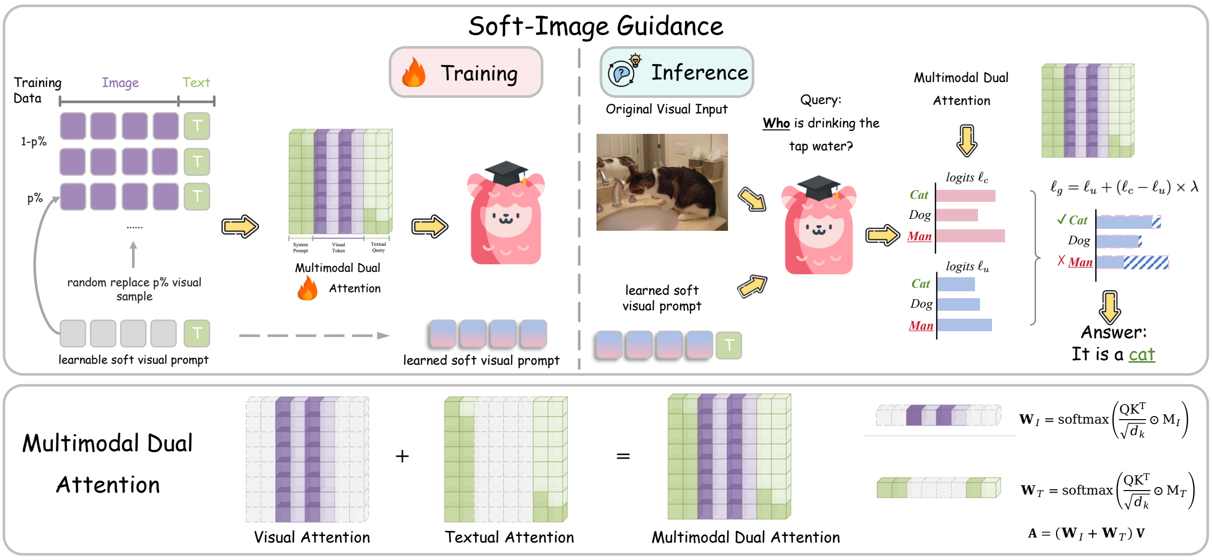

Figure 1: Overview of **LACING**, consisting of **Multimodal Dual Attention (bottom)** and **Soft-Image Guidance (above)** to mitigate language bias. MDA proposes a **parallel** dual-attention mechanism that constructs two **separate** attention for visual and text inputs. SIG implements a learnable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

context. As generation progresses, the model increasingly focuses on nearby generated text tokens while progressively neglecting fixed-position visual inputs (Zhang et al., 2024), as shown in Figure 4.

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 language bias of LVLMs from both 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.

To address training scale gaps in LVLMs, which leads to neglect of visual inputs across most layers (Chen et al., 2024a), we propose **Multimodal Dual-Attention Mechanism (MDA)**. Specifically, MDA introduces a parallel dual-attention mechanism that separately computes attention weights for each modality, and then fuses them to form the final attention map. This design ensures model to maintain substantial attention to visual inputs across all layers, promoting more effective visual-text integration. Crucially, unlike previous methods that apply bidirectional attention to visual inputs within a shared attention matrix (Xie et al., 2024; Zhou et al., 2024a), MDA builds parallel attention map that compute modality-specific attention scores separately. This separation enables flexible attention configurations; for instance, visual inputs can adopt either causal or bidirectional attention. In our design, we employ bidirectional attention for visual inputs to better capture global visual feature, while retaining causal attention for text to preserve the language modeling capabilities of LLMs.

To mitigate learned inference bias in LVLMs, we propose **Soft-Image Guidance (SIG)**, designed to enhance visual guidance by addressing the model's over-reliance on textual inputs (i.e., language bias). At core of SIG is a **learnable soft visual prompt**, which replaces visual inputs during both training and inference. It serves as a modality-aware placeholder, preserving input patterns (e.g., the input length and modalities), while implicitly compelling model to prioritize text inputs. Unlike prior methods (Leng et al., 2023; Zhang et al., 2024) that remove visual inputs or inject random noise, SIG maintains input consistency without introducing uncontrolled perturbations. During multimodal alignment stage, visual inputs are randomly replaced with soft prompt, allowing model to learn from complete and visual-substituted inputs. At inference, we replace visual inputs with well-learned soft prompt to form multimodal-null input. Each token's final output is computed by contrasting model's output distributions from original and multimodal-null inputs, ensuring each token in responses accounts for visual input more critically and thereby reducing language bias.

Our proposed MDA and SIG form a systematic framework for mitigating language bias in LVLMs, with each component complementing the other to further enhance overall performance. Comprehensive experiments across various model architectures and scales validate the effectiveness of LACING. We observe significant improvements, particularly in free-form generation and visual hallucinations reduction (e.g., 11.8-point gain on LLaVA-Bench (Liu et al., 2023) and a 40% improvement on Object Hall (Rohrbach et al., 2019; Yu et al., 2024)). Notably, LACING delivers consistent improvement without additional resource requirements beyond standard multimodal alignment setups (Liu et al., 2024c,d). Our analysis further confirms the efficacy of MDA in enabling LVLMs to fully utilize visual inputs, and robustness of SIG for reducing hallucinations and improving visual comprehension.

2 Related Work

2.1 Language Bias in LVLMs

Despite the impressive capabilities of LVLMs (OpenAI, 2024; Team, 2023; McKinzie et al., 2024; Wang et al., 2024a; Li et al., 2024), these models still struggle with generating responses irrelevant to the input images (Lan et al., 2024; Liu et al., 2024b), e.g., hallucinating non-existent objects (Zhou et al., 2024c). Zhao et al. (2024a) 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. (2024b) observe that LVLMs often answer questions using only LLM-derived textual knowledge. Chen et al. (2024a) further show that attention to visual inputs diminishes significantly in deeper layers, while Zhang et al. (2024) find that models increasingly prioritize text as generation progresses. These findings collectively indicate that LVLMs assign disproportionately low attention to visual inputs, limiting their ability to effectively utilize image information. Therefore, to address this challenge, we propose a systematic framework, LACING, that mitigates language bias from both training and inference perspectives.

2.2 Addressing Language Bias in LVLMs

Given the language bias of LVLMs, they exhibit similar hallucination issues as LLMs (Huang et al., 2023), as well as modality-specific hallucinations such as object hallucination (Rohrbach et al., 2019; Li et al., 2023c). As noted by Leng et al. (2023), this stems from the dominant influence of the LLM's pretraining distribution, making hallucination a prominent symptom of language bias. Recent efforts to mitigate hallucination fall into two main categories. The first includes training-intensive methods such as LRV (Liu et al., 2024a), LLaVA-BPO (Pi et al., 2024), LLaVA-RLHF (Sun et al., 2023), and RLHF-V (Yu et al., 2024), which rely on supervised fine-tuning or reinforcement learning with preference data. While effective, these methods typically necessitate substantial training data and computational resources. To address this, training-free methods have been proposed, including VCD (Leng et al., 2023), IBD (Zhu et al., 2024), VDD (Zhang et al., 2024), and ICD (Wang et al., 2024b). These methods contrast outputs with those from image-free inputs (or with distorted images) to reduce influence of textual LLMs. However, these methods may introduce inconsistencies between training and inference, limiting their effectiveness. Inspired by classifier-free guidance (Ho and Salimans, 2022), which combines conditional and unconditional signals for image generation, we propose a novel approach that addresses language bias from both training and inference perspectives and targets broader bias effects beyond object hallucination, improving general LVLM performance.

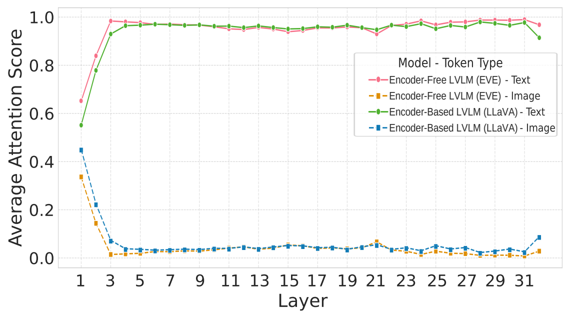

Figure 2: Average attention scores for output tokens towards text and visual tokens across different layers of encoder-based LVLMs (Liu et al., 2024c) and encoder-free LVLMs (Diao et al., 2024), showing that only the first two layers apply considerable attention to visual tokens. In contrast, deeper layers largely neglect them.

3 Method

3.1 Multimodal Dual-Attention Mechanism

Most LVLMs project bidirectional visual inputs into unidirectional LLM space using a relatively small amount of multimodal data (Liu et al., 2023, 2024c; Li et al., 2024) compared to vast pretraining data scales of LLMs (Dubey et al., 2024). LVLMs 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 data distribution changes using only shallow layers during training with limited data (Zhang et al., 2024). Consequently, LVLMs remains dominated by LLM's pretraining distribution and lacks effective attention to visual inputs in deeper layers. Shown in Figure 2, LVLMs (Bai et al., 2023; Wang et al., 2024a; Liu et al., 2024d; Diao et al., 2024) exhibit considerable attention toward visual inputs only in the first two layers (Chen et al., 2024a), while deeper layers retain their original distributions, causing deeper layers of LVLMs to ignore visual inputs. This pheromone has been observed across **various LVLMs**, including encoder-based LVLMs, such as LLAVA-Series (Liu et al., 2023, 2024c,d), QwenVL (Bai et al., 2023) and Qwen2VL (Wang et al., 2024a), and even encoder-free LVLMs like EVE (Diao et al., 2024) and Fuyu (Bavishi et al., 2023).

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 LLMs to allocate sufficient attention toward visual inputs and encourages LVLMs to fully leverage their LLM backbone for visual comprehension during training. This separation enables flexible attention configurations; for instance, visual inputs can adopt either causal or bidirectional attention. In our design, MDA retains causal attention for text inputs while independently calculating bidirectional attention towards visual inputs. As illustrated in Equation 1, given multimodal inputs $\mathbf{S} = \langle s_1, s_2, \ldots, s_N \rangle$, s_n means the token in inputs. To independently calculate attention weights across two modalities, we define two attention masks: mask $\mathbf{M}_{\mathcal{I}}$ for visual tokens $\mathcal{I}$ and mask $\mathbf{M}_{\mathcal{T}}$ for text tokens $\mathcal{T}$:

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

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

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

where $\mathbf{Q}, \mathbf{K}$ is query, key and in self-attention of LVLMs. Finally, the two attention weights $(\mathbf{W}_{\mathcal{I}})$ and $(\mathbf{W}_{\mathcal{T}})$, are integrated and multiplied by $\mathbf{V}$, the value in attention mechanism, to derive final attention score $\mathbf{A}$ based on MDA.

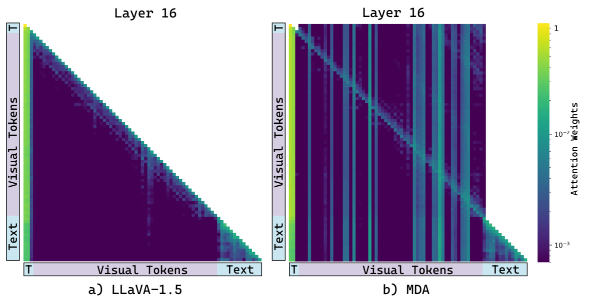

Figure 3: Attention allocation of a standard LVLM (LLaVA-1.5) and model trained with MDA. Text and visual tokens are marked in blue and purple, respectively.

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

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. 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. To support this design choice, we present a comparison between causal and bidirectional attention for visual inputs in § D.4.

3.2 Soft-Image Guidance

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.

Inspired by classifier-free guidance (Ho and Salimans, 2022) effectively combining the conditional and unconditional score to control the image generation quality, we propose the Soft-Image Guidance (SIG), designed to enhance the guidance of visual inputs during LVLMs' response generation and mitigate the inference bias of LVLMs. 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 Bayes' theorem, we have:

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

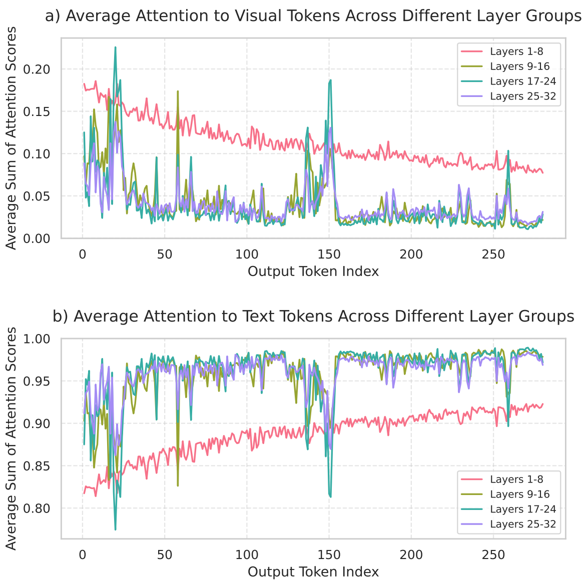

Figure 4: Attention allocation to visual and text tokens. Attention to visual tokens (a) decreases as response generates, while attention to text tokens (b) increases.

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

$$\log p(y_t \mid v) = \log p(v \mid y_t) + \log p(y_t) - \log p(v) \quad (5)$$

In Eq. (5), $p(y_t)$ is unconditional probability of generating token y_t without visual input.

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

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

To express $\log \widehat{p}(y_t \mid v)$ with known quantities, we expand $\log p(v \mid y_t)$ using Bayes' theorem:

$$\log p(v \mid y_t) = \log p(y_t \mid v) + \log p(v) - \log p(y_t) \quad (7)$$

Substituting Eq. (7) into Eq. (6), we obtain:

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

Since v is given (fixed), $\log p(v)$ is constant for $y_t$and can be omitted, we simplify Eq. (8) to:

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

Eq. (9) demonstrates that influence of visual input v on text generation can be amplified by adjusting scaling parameter λ, once given conditional probability $p(y_t \mid v)$ of original inputs and unconditional probability $p(y_t)$ without visual inputs. This formulation highlights a major challenge in enhancing visual guidance for LVLMs: accurately calculating unconditional probability $p(y_t)$ of generating token y_t in the absence of visual input.

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** θ ▷ Replace visual input with soft prompt ϵ
  $\mathcal{L}_{\text{cross-entropy}} = -\mathbb{E}_{(\mathbf{x},\mathbf{v})} \sum_i \mathbf{y}_i \log P(\mathbf{x}, \mathbf{v})$
  Update P and ϵ
until converged

Previous approaches attempt to ascertain such probabilities probability by either providing the model with text-only input (Zhang et al., 2024) or by injecting randomly generated noise to mask the image (Leng et al., 2023), 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 (Chen et al., 2024a; Zhang et al., 2024). 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 $p(y_t)$.

SIG first employs a learnable soft visual prompt ϵ to replace the visual input, 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 a dual purpose, acting both as a placeholder to maintain the input pattern and as an indicator to make the model prioritize text input. 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 the 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:

$$\ell_g = \ell_u + (\ell_c - \ell_u) \times \lambda \quad (10)$$

Model	Model Size	Obj Hall		MMHall		LLaVABench↑	MM-VET↑
		Res ↓	Obj ↓	Score ↑	Hall ↓		
LRV† (Liu et al., 2024a)	7B	32.30	22.30	/	/	/	31.70
LLaVA-1.5† (Liu et al., 2024c)	7B	46.71	25.08	2.19	59.00	64.40	31.10
VCD‡ (Leng et al., 2023)	7B	47.40	25.24	2.12	59.00	65.30	30.90
VDD-None‡ (Zhang et al., 2024)	7B	46.71	25.19	2.22	56.00	66.00	31.70
ICD‡ (Wang et al., 2024b)	7B	47.40	25.00	2.18	59.00	64.70	31.10
Less-is-more‡ (Yue et al., 2024)	7B	40.30	17.80	2.33	50.00	60.90	/
OPERA‡ (Huang et al., 2024)	7B	45.10	22.30	2.15	54.20	60.30	/
HA-DPO° (Zhao et al., 2024b)	7B	39.90	19.90	1.98	60.40	67.20	/
POVID° (Zhou et al., 2024b)	7B	48.10	24.40	2.08	56.20	62.20	/
LLaVA1.5-7B-BPO° (Pi et al., 2024)	7B	31.90	15.10	/	/	71.60	**36.80**
LACING	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† (Liu et al., 2023)	13B	63.00	29.50	/	/	70.80	26.40
Muffin† (Lou et al., 2024)	13B	50.50	24.50	/	/	68.80	/
QWEN-VL† (Bai et al., 2023)	10B	40.40	20.70	/	/	52.10	/
LLaVA-1.5† (Liu et al., 2024c)	13B	47.06	23.33	2.54	50.00	72.50	36.10
VCD‡ (Leng et al., 2023)	13B	46.37	23.10	2.60	49.00	73.60	36.90
VDD-None‡ (Zhang et al., 2024)	13B	44.64	22.23	2.38	55.00	73.00	36.10
ICD‡ (Wang et al., 2024b)	13B	45.52	21.93	2.41	54.00	72.50	36.20
LLaVA-RLHF° (Sun et al., 2023)	13B	38.10	18.90	2.53	57.00	61.50	/
RLHF-V° (Yu et al., 2024)	13B	**12.20**	**7.50**	2.45	51.00	51.40	/
LLaVA1.5-13B-BPO° (Pi et al., 2024)	13B	27.30	12.90	/	/	74.40	**41.40**
LACING	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 across multiple benchmarks, highlighting highest score in **bold** and second highest underlined. Baselines are categorized as: † (LVLMs), ‡ (training-free), and ○ (reinforcement learning-based).

Eq. (10) facilitates a more balanced and effective integration of visual inputs, enhancing visual comprehension while addressing the language bias.

4 Experiments

4.1 Implementation Details

To ensure fair comparison and validate the effectiveness of our approach, we train LVLMs from scratch and evaluate against strong baselines. Given availability of open-sourced multimodal alignment datasets, we select two representative LVLMs with different architectures and model scales: LLaVA-1.5 (Liu et al., 2024c) and LLaVA-Next (Liu et al., 2024d) as our base model. We strictly follow their training settings, including the same dataset and model backbone. The model is trained on 8 A100 GPUs, each with 40 GB of memory. Details of scaling parameter λ and replacement probability θ are shown in § B.3. Additional information, including extra costs discussion, training and experiment details, can be found in § D, § B.1, § B, and § C.

4.2 Evaluation Setup

We conduct experiments across three categories:
Visual Comprehension: MMBench(Liu et al., 2024e) evaluates fine-grained abilities of LVLMs, assessed with accuracy. **TextVQA** (Singh et al., 2019) employs VQA accuracy (Agrawal et al., 2016) as metric for questions with text within images. We send models with pure images for evaluation. **MM-VET** (Yu et al., 2023) evaluates LVLMs with GPT-4 in free-form question-answering.
Open-ended Generation: LLaVA-Bench (Liu et al., 2023) uses GPT-4 to compare generated answers with reference answers.
Visual Hallucination: MMHal-Bench (Sun et al., 2023) evaluates hallucinations and response informativeness, with GPT-4 comparing model outputs to human responses and object labels. **Object Hall-Bench** (Rohrbach et al., 2019) detects object hallucinations by comparing model outputs with COCO labels (Lin et al., 2015). We follow same setup as (Yu et al., 2024), which adds diverse prompts with detailed image descriptions for evaluations.

4.3 Experimental Results

We evaluate our method across benchmarks in Table 2, comparing with baseline models: (1) LVLMs after multimodal alignment training(†); (2) training-free methods for mitigating hallucinations(‡); and (3) reinforcement learning methods(○). LACING consistently outperforms across all benchmarks. Notably, over LLaVA-1.5 (Liu et al., 2024c), which shares same training data and architecture, LACING achieves double-digit percentage gains across different model sizes(indicated by Δ), demonstrating strong scalability. LACING also surpasses

Method	Model Size	MMBench↑	TextVQA↑	LLaVABench↑	Obj Hall	
					Res ↓	Obj ↓
Greedy Sampling						
LLaVA-1.5	7B	64.61	46.05	64.40	46.71	25.08
VCD	7B	64.69 (+ 0.08)	46.05 (+ 0.00)	65.30 (+ 0.90)	47.40 (+ 0.69)	25.24 (+ 0.16)
VDD-None	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)
Nucleus Sampling						
LLaVA-1.5	7B	56.96	35.41	63.00	56.66	29.75
VCD	7B	60.91 (+ 3.95)	40.67 (+ 5.26)	65.30 (+ 2.30)	49.83 (- 6.83)	27.44 (- 2.31)
VDD-None	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)

Table 2: Comparison of SIG with training-free methods designed to mitigate hallucinations under various decoding strategies. Performance gap compared to the base model(LLaVA-1.5) are noted in parentheses. Red denotes improvements, ; green indicates negative effects. Additional results for other model sizes are in § D.2.

training-free methods such as VCD (Leng et al., 2023), VDD (Alabdulmohsin et al., 2024) and ICD (Wang et al., 2024b), achieving nearly 20 points reduction on Obj Hall. The underperformance of these methods further indicates that adding randomly generated noise on input images or simply remove images during the inference injects the unexpected information that was not present during training, thereby diminishing robustness of their methods. Compared to reinforcement learning-based methods, which require extensive training resources and additional high-quality feedback data, LACING remains effective and cost-efficient while delivering superior results. While RLHF-V achieves best score on Obj Hall, likely due to overfitting from overlap with its training data, base model, and benchmark (Yu et al., 2024; Lou et al., 2024). In contrast, LACING outperforms RLHF-V by a wide margin in other tasks (e.g., +32.9 on LLaVABench). Overall, our model demonstrates lower hallucination rates and higher visual comprehension scores without requiring additional resources, showcasing the effectiveness of our proposed method. For thorough evaluations, we conduct experiments across various benchmarks in § D.1, including ScienceQA (Lu et al., 2022), POPE (Li et al., 2023c), SeedBench (Li et al., 2023a), and MMVP (Tong et al., 2024), showing consistent improvements. We also perform LACING on LLaVA-Next to demonstrate the generalization across different model architectures in § D.3.

4.4 Analysis Results

Effect of SIG in Decoding Perspective To distinguish LACING from prior works, we investigate effectiveness of SIG in different decoding strategies. As shown in Table 2, existing training-free methods, like VCD (Leng et al., 2023) and VDD-None (Zhang et al., 2024), only yield gains under **Nucleus Sampling** (Holtzman et al., 2020), while SIG consistently improves performance under both **Greedy** and **Nucleus Sampling**. It is further validated across different model sizes in § D.2. VCD contrasts outputs from original and distorted visual inputs, while VDD uses text-only inputs. However, Adding random noise or omitting visual inputs at inference create discrepancies not present during training, leading to degraded performance and reduced robustness, especially on benchmarks like MMBench, where outputs are short and deterministic. Greedy Sampling, which selects most probable token, offers limited tolerance for the introduced noise, making these methods less effective. By contrast, Nucleus Sampling introduces randomness by sampling from a probability distribution, which mitigate sensitivity to noise, making these methods appear effective. However, this randomness may harm performance in tasks requiring precise outputs (e.g., multi-choice QA), often underperforming compared to Greedy Sampling. In contrast, SIG replaces visual inputs with a learnable soft visual prompt that preserves input patterns while compelling model to prioritize text inputs. It ensures consistency between training and inference, enabling SIG to deliver robust gains under both decoding strategies. Additional comparisons in § D.2 further demonstrate SIG's effectiveness against IBD (Zhu et al., 2024), ICD (Wang et al., 2024b), VDD-UNK (Zhang et al., 2024), and a variant using a blank image.

How do LVLMs Treat Visual Inputs with MDA? To evaluate the effectiveness of MDA in mitigating language bias caused by training scale disparities, we analyze how LVLMs process visual inputs

Model	LLaVABench			
	Complex	Conv	Detail	All
LLaVA-1.5	75.50	54.10	56.60	64.40
w. FastV	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	10.70	10.20	10.40	10.50
Δ	**- 72.50**	**- 49.50**	**- 48.80**	**- 59.80**

Table 3: Performance on LLavaBench between LLaVA-1.5 and those with MDA, with and without FastV.

Sampling	Model	TextVQA	LLavaBench	MM-VET
Greedy	LLaVA-1.5	46.05	64.40	31.10
	LACING	**46.94**	**72.20**	**33.50**
	-w/o. MDA	46.77	70.60	32.00
	-w/o. SIG	46.03	70.30	32.80
Nucleus	LLaVA-1.5	35.41	63.00	29.80
	LACING	**42.05**	**72.20**	**35.20**
	-w/o. MDA	39.40	68.40	33.30
	-w/o. SIG	36.40	67.80	30.50

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

across layers. To assess whether MDA addresses this issue, we adopt the pruning method proposed by Chen et al. (2024a) on LLaVA-1.5 with MDA by pruning half of the visual tokens in deeper layers and measuring performance on LLaVA-Bench. Prior work (Chen et al., 2024a) shows that pruning visual tokens in deeper layers has minimal impact on standard LVLMs, indicating poor utilization of visual inputs at those layers. In contrast, our results in Table 3 show a significant performance drop when pruning is applied to the model with MDA, confirming that visual information is effectively utilized throughout all layers—not just shallow ones. MDA ensures comprehensive attention to visual inputs across the model's layers, thereby facilitating LVLMs in fully exploiting its visual comprehension capabilities. The 7.7-points improvement for complex tasks on LLaVABench in Table 3 validate this conclusion, as complex tasks generally require deeper layers for precise understanding (Ben-Artzy and Schwartz, 2024; Jin et al., 2024).

Ablation Study To understand contributions of each component, we conduct an ablation study across multiple benchmarks in Table 4 on the 7B model under different decoding strategies. 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. 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. To further validate LACING, we conduct ablation studies across various model sizes on multiple benchmarks in § D.5.

Effectiveness on Different Model Architecture To validate robustness of LACING, we conduct additional experiments on other model architectures. We use LLaVA-NEXT (Liu et al., 2024d) as base model, which supports dynamic resolution. Due to training data availability, we leverage training data from fully open-sourced version of LLaVA-NEXT (Chen and Xing, 2024). Results show that our approach applies to LLaVA-NEXT as well, proving its versatility across different architectures and training methods. See § D.3 for details.

Effect of Bidirectional Attention in MDA for Visual Inputs. To validate our design choice and highlight that the core strength of MDA lies in its parallel dual-attention mechanism, we compare attention strategies for visual inputs in § D.4. Results show that even with causal attention, MDA outperforms the baseline, confirming the effectiveness of the dual-attention design. Bidirectional attention yields greater gains, aligning better with the spatial nature of visual data and justifying its use in MDA.

Parameter Study. We conduct the parameter study in § B.3 with the detailed discussion.

Human Evaluation and Case Study. The human evaluation on LLaVABench and a practical case study are detailed in § G and § H, respectively, demonstrating effectiveness of LACING.

5 Conclusion

This paper tackles the language bias in LVLMs, which often leads to neglect of visual inputs and hallucinatory responses. We identify two primary sources of this bias: gap in training scales between the pretraining and multimodal alignment, and learned inference bias. To reduce language bias, we introduced 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 advanced LVLMs.

6 Limitation

Despite the promising results demonstrated by LACING in addressing the language bias of LVLMs, several limitations must be acknowledged. First, although we validate the effectiveness of our method on two representative LVLMs that has different architecture—LLaVA-1.5 and LLaVA-Next—more extensive evaluation across a wider range of LVLM architectures is still lacking. This is primarily because our method targets the multimodal alignment stage that post-trains an LLM-based backbone into an LVLM, requiring fair comparisons that retrain models from scratch. However, for more advanced LVLMs such as Qwen-VL-2.5 and InternVL-3, the data and training details for their multimodal alignment stages are not fully open-sourced, making it infeasible to apply or evaluate our approach directly. Nevertheless, language bias is commonly observed across various LVLMs (Zhao et al., 2024a; Chen et al., 2024b,a) and even the SOTA LVLMs (Wang et al., 2024a) exhibits such phenomena. Therefore, inspired by this common observation and the consistent gains observed across model sizes and different in our experiments, 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.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Ibrahim Alabdulmohsin, Vinh Q. Tran, and Mostafa Dehghani. 2024. Fractal patterns may illuminate the success of next-token prediction. *Preprint*, arXiv:2402.01825.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Amit Ben-Artzy and Roy Schwartz. 2024. Attend first, consolidate later: On the importance of attention in different llm layers. *Preprint*, arXiv:2409.03621.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *Preprint*, arXiv:2403.06764.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are we on the right way for evaluating large vision-language models? *Preprint*, arXiv:2403.20330.

Lin Chen and Long Xing. 2024. Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. `https://github.com/xiaoachen98/Open-LLaVA-NeXT`.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Michał Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. 2017. Frustratingly short attention spans in neural language modeling. *Preprint*, arXiv:1702.04521.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. 2024. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*.

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

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *Preprint*, arXiv:2207.12598.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

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

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. 2024. Exploring concept depth: How large language models acquire knowledge at different layers? *Preprint*, arXiv:2404.07066.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. 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.

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

Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. 2024. A survey of hallucination in large visual language models. *Preprint*, arXiv:2410.15359.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. 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. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *Preprint*, arXiv:2306.14565.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

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. 2024e. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. 2024. Muffin: Curating multi-faceted instructions for improving instruction-following. *Preprint*, arXiv:2312.02436.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

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

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. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *Preprint*, arXiv:2403.09611.

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

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. hello-gpt-4o.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *Preprint*, arXiv:2403.08730.

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. 2021. 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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

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

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

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

ShareGPT. 2023. `https://sharegpt.com/`.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. 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.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

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. 2023. Aligning large multimodal models with factually augmented rlhf. *Preprint*, arXiv:2309.14525.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv: 2312.11805*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.

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

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *Preprint*, arXiv:2403.18715.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *Preprint*, arXiv:2408.12528.

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

Ruiqing Yan, Linghan Zheng, Xingbo Du, Han Zou, Yufeng Guo, and Jianfei Yang. 2024. Recurformer: Not all transformer heads need self-attention. *Preprint*, arXiv:2410.12850.

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

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. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *Preprint*, arXiv:2312.00849.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Y. X. Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. 2025. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *Preprint*, arXiv:2502.11089.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *Preprint*, arXiv:2402.14545.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2024. Debiasing multimodal large language models. *Preprint*, arXiv:2403.05262.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2024a. Mmicl: Empowering vision-language model with multi-modal in-context learning. *Preprint*, arXiv:2309.07915.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2024b. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *Preprint*, arXiv:2311.16839.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024a. Transfusion: Predict the next token and diffuse images with one multi-modal model. *Preprint*, arXiv:2408.11039.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024b. Aligning modalities in vision large language models via preference fine-tuning. *Preprint*, arXiv:2402.11411.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024c. Analyzing and mitigating object hallucination in large vision-language models. *Preprint*, arXiv:2310.00754.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *Preprint*, arXiv:2402.18476.

A Appendix

This Appendix is organized as follows.

- In § B, we show implementation details of our method: training details(§ B.1), datasets(§ B.2) and hyperparameters(§ B.3).
- In § C, we present the details of our experiments and evaluation. Specifically, dataset and metric(§ C.1), baselines(§ C.2) and GPT-4 Version(§ C.3)
- In § D, we provide the additional experiments, including the evaluations across wide-range of benchmarks(§ D.1), baselines(§ D.2), different architecture(§ D.3), different design choice(§ D.4) and different model size(§ D.5).
- In § E, we analyze early-fusion LVLMs like Chameleon, trained from scratch with a balanced mix of text and visual tokens, distinguishing them from the LVLMs discussed in this paper.
- In § F, we detail the experiments and provide an in-depth discussion on the impact of hyperparameters, specifically the replace probability θ(§ F.1) and the scaling parameter λ(§ F.2).
- In § G, we present a human evaluation of LACING versus LLaVA-1.5 across LLaVABench.
- In § H, we present more qualitative results.
- In § I, we visualized the attention distribution across different layers in LLaVA-1.5 and LACING.

B Training Details

To make fair compression, we adopt the same training settings as LLaVA-1.5 (Liu et al., 2024c), 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.

B.1 Training

Following the setting of LLaVA-1.5 (Liu et al., 2024c), we employ CLIP-ViT-L-14-336 (Radford et al., 2021) 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 (Chiang et al., 2023) 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 (Rajbhandari et al., 2020) and Deepspeed Zero3 (Rajbhandari et al., 2020) 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.

Dataset	Data Size
LLaVA (Liu et al., 2023)	158K
ShareGPT (ShareGPT, 2023)	40K
VQAv2 (Goyal et al., 2017)	83K
GQA (Hudson and Manning, 2019)	72K
OKVQA (Marino et al., 2019)	9K
OCRVQA (Mishra et al., 2019)	80K
A-OKVQA (Schwenk et al., 2022)	66K
TextCaps (Sidorov et al., 2020)	22K
RefCOCO (Kazemzadeh et al., 2014)	48K
VG (Krishna et al., 2017)	86K
Total	665K

Table 5: Instruction-following Data Mixture Used for Finetuning (Liu et al., 2024c).

B.2 Data

We strictly follows the data setting of LLaVA-1.5 for both pretraining and finetuning. Specifically, the LLaVA-558K (Liu et al., 2023) for pertraining and a mixture of instruction-following data for

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 (Liu et al., 2024c), except that we set the replace probability θ for training with soft-image guidance.

finetuning shown in Table 5.

B.3 Hyperparameters

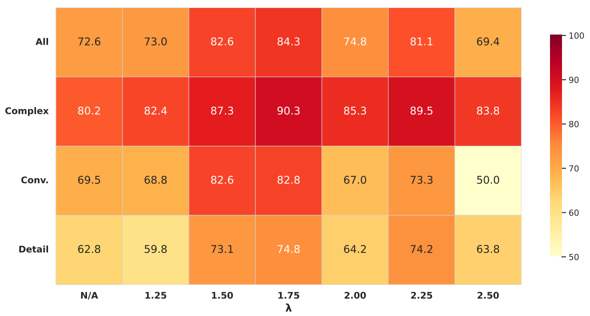

Figure 5: Model performance on LLaVABench across various scaling parameter λ.

We utilize the identical set of hyperparameters as the original LLaVA-1.5 (Liu et al., 2024c), 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 (Liu et al., 2024e) 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 (Liu et al., 2024e) 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 (Singh et al., 2019) is designed for visual question answering involving text within images. It employs VQA accuracy as the evaluation metric. Unlike LLaVA-1.5 (Liu et al., 2024c), 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 (Yu et al., 2023) 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 (Liu et al., 2023) 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 (Rohrbach et al., 2019) detects object hallucinations by comparing model outputs with COCO image labels (Lin et al., 2015). Yu et al. (2024) 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.

MMHall-Bench (Sun et al., 2023) 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.

Model	SCIQA↑	POPE↑			SeedBench↑			MMVP↑
LLaVA-1.5	70.12	87.38	84.26	86.21	58.60	66.10	37.30	26.00
VCD	70.12	87.39	84.25	86.21	59.93	65.62	38.41	26.00
LACING	**71.26**	**87.74**	**85.60**	**86.50**	**61.35**	**67.46**	**38.19**	**32.00**

Table 7: Experiments with more benchmarks across 7B model

Model	MMBench	TextVQA
Greedy Sampling		
LLaVA-1.5 (Liu et al., 2024c)	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 (Liu et al., 2024c)	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 8: Performance comparison of models undergoes training for one or two epochs across MMBench and TextVQA.

C.2 Baselines

General LVLMs that have undergone multi-modal alignment training. Specifically, we utilize LLaVA (Liu et al., 2023), Qwen VL (Bai et al., 2023), LLaVA-1.5 (Liu et al., 2024c), Muffin (Lou et al., 2024), and LRV (Liu et al., 2024a) as representative baselines. These LVLMs are predominantly trained with multimodal data for alignment (Liu et al., 2023; Bai et al., 2023; Lou et al., 2024) and fine-tuned using high-quality instruction data (Liu et al., 2024c,a), thereby achieving exceptional performance in various multimodal tasks. For example, LRV (Liu et al., 2024a) 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 (Leng et al., 2023) contrast model outputs generated from original inputs and distorted visual input to reduce over-reliance on statistical bias and unimodal priors. Similarly, VDD (Zhang et al., 2024) contrast model outputs from original inputs and inputs without visual inputs to reduce the influence of textual LLMs. OPERA (Huang et al., 2024) 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 (Yue et al., 2024) 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 (Zhou et al., 2024b) 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 (Rafailov et al., 2024) to mitigate hallucinations. HA-DPO (Zhao et al., 2024b) propose a pipeline for constructing positive and negative sample pairs and adopt the direct preference optimization (Rafailov et al., 2024) using the constructed dataset to reduces hallucination. RLHF-V (Yu et al., 2024) employs the Muffin (Lou et al., 2024) 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 (Pi et al., 2024) proposes a pipeline to gather preference datasets and conduct preference learning to mitigate this type of hallucination.

C.3 GPT-4 Version

For all evaluations conducted using the GPT-4(evaluation for Object HalBench, MMHall-Bench, LLaVABench, and MM-VET), we utilized the `GPT-4 API` in October 2024. It ensures consistency with prior research (Yu et al., 2023, 2024; Sun et al., 2023; Liu et al., 2023). According to the documentation provided by OpenAI[1], `GPT-4 API` currently points to `GPT-4-0613 API`.

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

Method	LLaVABench↑	MM-VET↑	MMHall		Obj Hall	
			Score ↑	Hall ↓	Res ↓	Obj ↓
LLaVA-1.5 (Liu et al., 2024c)	64.40	31.10	2.19	59	46.71	25.08
IBD (Zhu et al., 2024)	64.60	31.10	2.24	58	46.31	24.16
ICD (Wang et al., 2024b)	64.70	31.10	2.18	59	47.40	25.00
VDD-UNK (Zhang et al., 2024)	65.30	31.00	2.22	56	46.71	24.82
SIG-blank	68.40	31.50	2.42	52	34.41	17.80
SIG	**70.60**	**32.00**	**2.47**	**50**	**30.36**	**15.16**

Table 9: Comparison of SIG with other baselines on 7B model

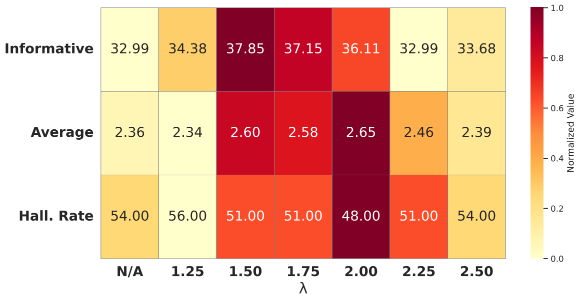

Figure 6: Model performance on MMHall-Bench across various scaling parameter λ.

D Additional Experiments

D.1 Additional Evaluations across other benchmarks

To further demonstrate the generalizability of LACING, we conducted experiments on additional benchmarks, including ScienceQA, POPE, Seed-Bench, and MMVP. The results presented in Table 7 consistently show improvements, confirming the effectiveness of our method.

D.2 Comparison Between SIG and Other Methods

Table 9 compares SIG with other training-free baselines, including a variant using a blank image instead of the learnable soft-image prompt. The results show that SIG outperforms all baselines, with the learnable prompt significantly surpassing the blank-image variant while adding only 0.02–0.03% more parameters.

Table 10 compares SIG with other training-free baselines for the 13B model. The results confirm that while prior training-free approaches improve performance only with Nucleus Sampling, SIG demonstrates effectiveness across all decoding settings.

D.3 Evaluation Across Different Model Architectures

To ensure a fair comparison, we train the LVLM from scratch using our method and evaluate its performance against baseline models. Given the availability of training data, we select LLaVA-1.5 (Liu et al., 2024c) as our base model and strictly adhere to its training settings, including the same dataset and model backbone. To further validate the robustness of our approach, we conduct additional experiments across various model architectures. Specifically, we use LLaVA-NEXT (Liu et al., 2024d) as the base model, which supports dynamic resolution. Due to training data availability, we leverage the dataset from the fully open-sourced version of LLaVA-NEXT (Chen and Xing, 2024) and adhere to its training settings. We set the Vicuna-1.5 (Chiang et al., 2023) language model backbone and ViT-L-14-336 (Radford et al., 2021) as the visual encoder. Our preliminary results, presented in Table 11, indicate that similar performance trends hold across additional LVLMs. This underscores that our approach is not limited to a specific architecture or training setup.

D.4 Comparison of Different Attention Mechanism for Visual Inputs in MDA

To validate our design choice and highlight that the core strength of MDA lies in its parallel dual-attention mechanism, we compare different attention strategies for visual inputs in Table 12. The results show that even when using only causal attention, MDA still yields performance gains over the baseline, confirming the effectiveness of the dual-attention design. However, bidirectional attention achieves more significant improvements, aligning more naturally with the spatial characteristics of visual data. This further supports our motivation for adopting bidirectional attention for visual inputs in MDA.

Method	Model Size	MMBench↑	TextVQA↑	LLaVABench↑	Obj Hall	
					Res ↓	Obj ↓
Greedy Sampling						
LLaVA-1.5	13B	67.74	48.66	72.50	47.06	23.33
VCD	13B	68.38 (+ 0.64)	48.63 (- 0.03)	73.60 (+ 1.10)	46.37 (- 0.69)	23.10 (- 0.23)
VDD-None	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	13B	62.11	38.92	68.10	50.52	25.74
VCD	13B	65.38 (+ 3.27)	43.56 (+ 4.64)	70.70 (+ 2.60)	49.83 (- 0.69)	24.23 (- 1.51)
VDD-None	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 10: Comparison of SIG with training-free methods under different decoding strategies in 13B model. Performance gap compared to the base model(LLaVA-1.5) are noted in parentheses. Red denotes positive improvements, while green indicates negative effects.

Model	Obj Hall		MMHall		MM-VET ↑
	Res ↓	Obj ↓	Score ↑	Hall ↓	
LLaVA-Next	13.81	7.50	2.67	51.00	37.6
LACING	**7.92**	**4.29**	**2.84**	**49.00**	**42.2**

Table 11: Performance of LACING on LLaVA-Next.

Method	**MM-VET ↑**	**LLavaBench ↑**
LLaVA-1.5	31.10	64.40
Causal Attn.	31.90	69.60
Bi-Attn.(MDA)	**32.80**	**70.30**

Table 12: Comparison of different visual attention strategies in MDA.

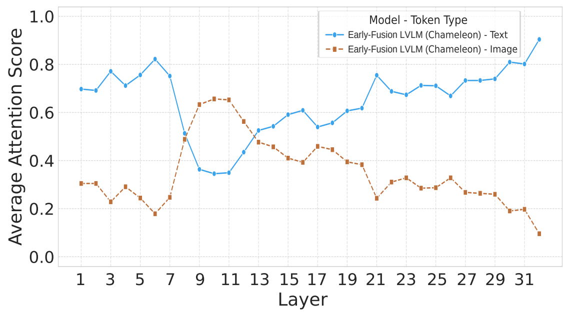

Figure 7: Average attention scores for output tokens towards text and visual tokens across different layers of early-fusion LVLMs (Chameleon (Team, 2024)).

D.5 Ablation Studies Across Different Model Size

To further validate our method, we conduct ablation studies across various model sizes on multiple benchmarks. Specifically, we perform an ablation study on the 13B model across multiple benchmarks to analyze the impact of different components. Table 13 presents the results, demonstrating that our approach outperforms the baseline and its ablated variants across both MMBench and LLaVABench, under both greedy decoding and sampling strategies.

	MMBench		LLaVABench	
	Greedy	Sampling	Greedy	Sampling
LLaVA-1.5	67.74	62.11	72.5	68.1
w.o. SIG	68.73	65.55	76.7	75.5
w.o. MDA	68.99	64.77	74.7	72.0
LACING	**70.01**	**66.92**	**78.5**	**76.6**

Table 13: Ablation study on 13B models.

E Analysis of Early-fusion LVLMs

The performance of LVLMs is often hindered by the disparity in training scales between the LLM pretraining phase and the subsequent LVLM alignment stage. This imbalance results in suboptimal utilization of visual inputs, as evidenced by the attention distributions: only the initial layers demonstrate significant attention to visual tokens, while the deeper layers tend to neglect them. In contrast, early-fusion LVLMs, such as Chameleon (Team, 2024), which are trained from scratch using a balanced mix of visual and textual tokens, exhibit a more effective modality fusion. As shown in Figure 7, this balanced training approach enables the model to allocate attention more uniformly across modalities, thereby mitigating the issues associated with scale disparities during pretraining and alignment.

Following pervious work (Zhao et al., 2024a), we measure performance gaps on image-required vs.non-image-required questions gathered from Science QA (Lu et al., 2022) dataset to evaluate language bias. As shown in Table 14, although showing better fusion, Chameleon, as well as other LVLMs still remains substantial language bias.

F Parameter Study

Model	Don't Req	Req	Gap
LLaVA	56.78	72.84	16.06
EVE	68.13	45.33	22.80
Chameleon	56.12	37.33	18.79

Table 14: Language Bias Evaluation.

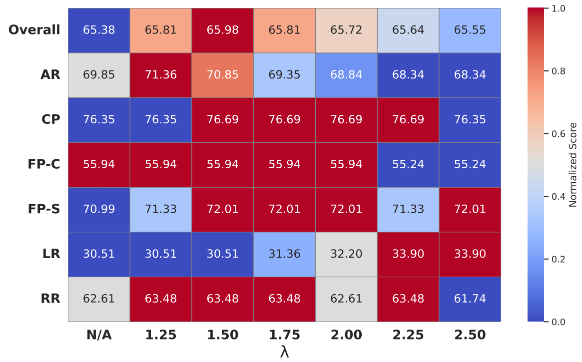

Figure 8: Model performance on MMBench across various scaling parameter λ.

F.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 8 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 15 to identify the optimal value for this parameter.

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

Table 15: Performance of SIG on MMBench and LLavaBench across different replace probability θ

F.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 8, 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.

G 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 (Liu et al., 2024c). 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 9 showcases the effectiveness of our method.

During the human evaluation, the participants

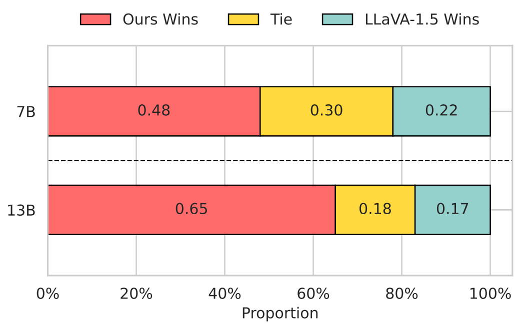

Figure 9: Human evaluation on LLaVABench.

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.
3. **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. The human evaluation results in Figure 9 shows that LACING can generate responses that consistently outperformed baseline models across all three evaluation criteria. These results highlight the model's ability to deliver high-quality answers that are both factually accurate and contextually relevant, while maintaining fluency and coherence.

H 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 utilizing samples from LLaVABench. The evaluations of the case studies on the 13B model are illustrated in Figure 10, Figure 11, Figure 12, and Figure 13. Case studies demonstrate that LACING effectively reduces hallucinations while generating accurate responses, showcasing fine-grained visual comprehension abilities.

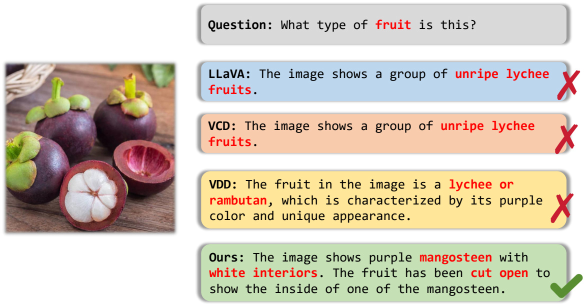

Figure 10: Comparison of LACING with other baselines on a sample from LLaVABench. Hallucinated responses and our corrections are highlighted in **red**.

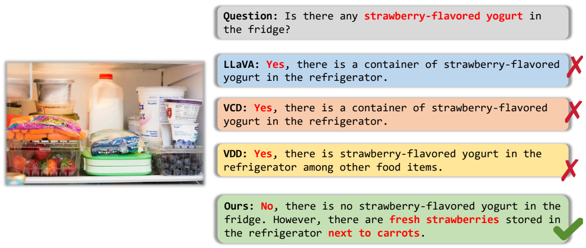

Figure 11: Comparison of LACING with other baselines on a sample from LLaVABench. LACING demonstrates a reduction in object hallucination and an enhancement in fine-grained visual comprehension, such as the identification of fresh strawberries in the refrigerator.

I 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 14 and Figure 15, comparing the standard LVLM (LLaVA-1.5) with MDA. We get the attention allocation of LVLMs shown in this paper, such as Figure 14, and Figure 15, 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 of the main paper.

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 12: 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 13: 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.

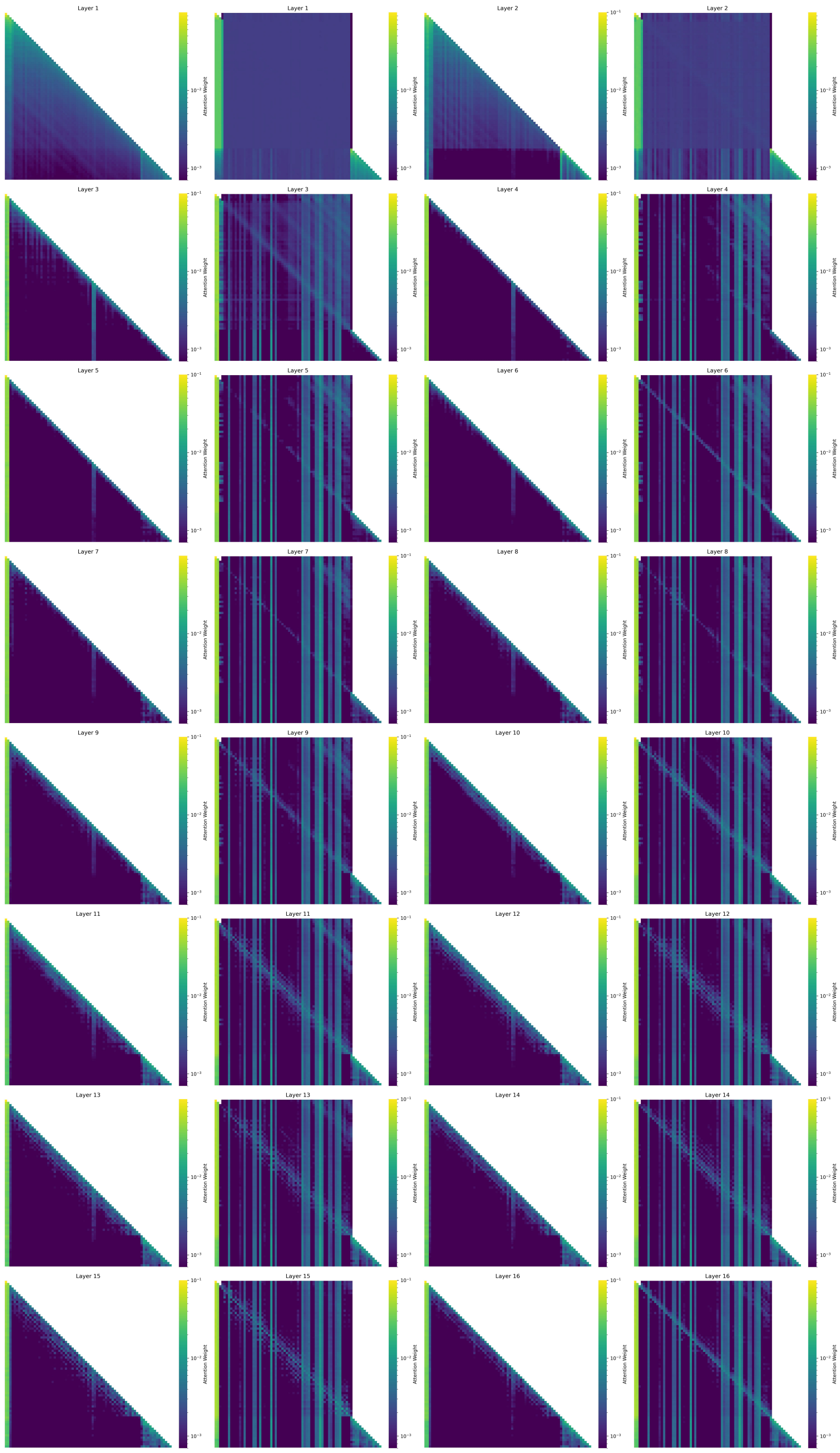

Figure 14: Comparison of Attention Maps across the 1st to 16th Layer in LLaVA and LACING.

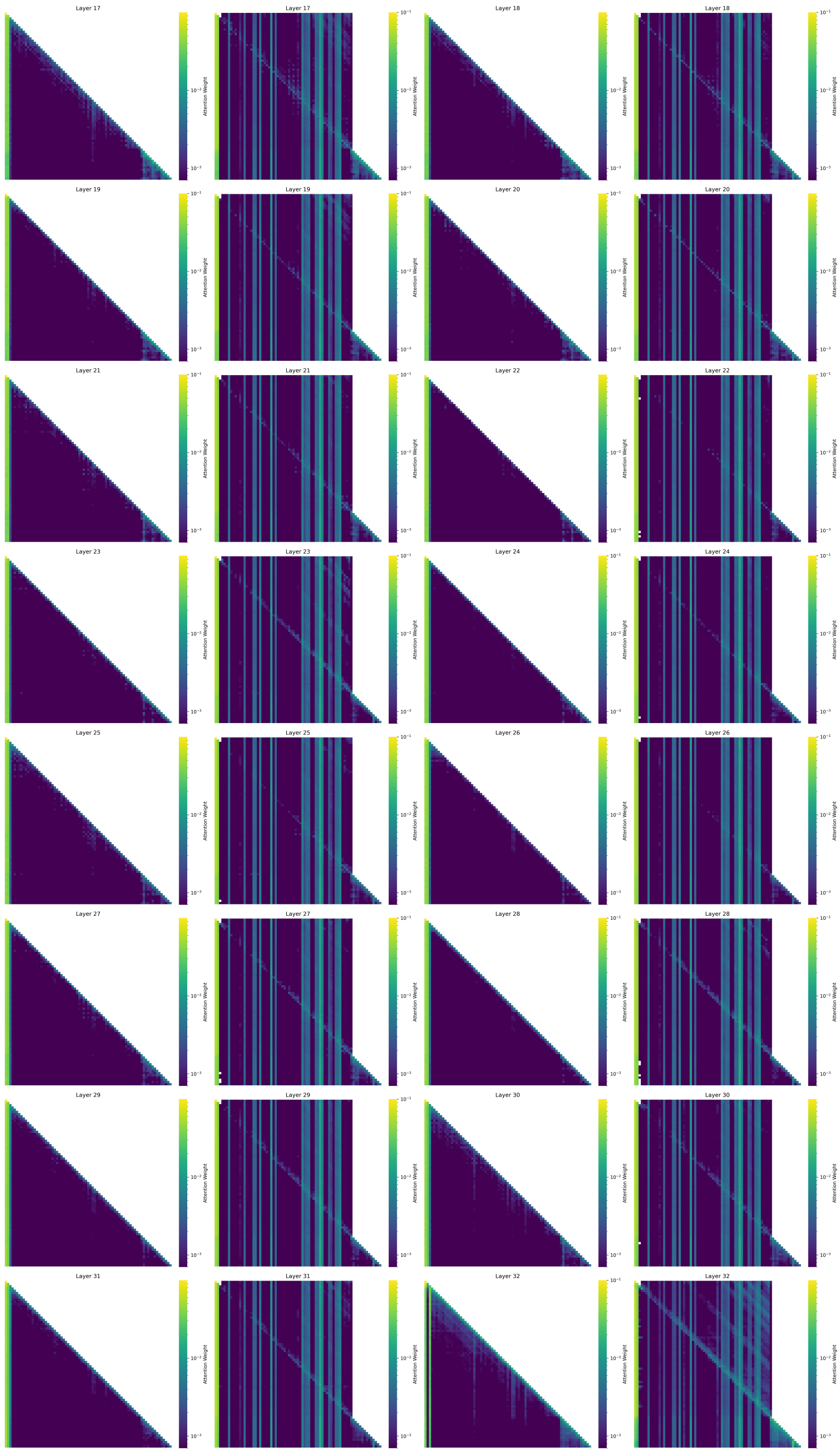

Figure 15: Comparison of Attention Maps across the 17th to 32nd Layer in LLaVA and LACING.