Grounding Language with Vision: A Conditional Mutual Information Calibrated Decoding Strategy for Reducing Hallucinations in LVLMs

Hao Fang[*1], **Changle Zhou**[*1], **Jiawei Kong**[*1], **Kuofeng Gao**[1],
Bin Chen[2], **Tao Liang**[1], **Guojun Ma**[†1], **Shu-Tao Xia**[†1]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University,
[2]Harbin Institute of Technology, Shenzhen,
{fangh25, kjw25, gkf21}@mails.tsinghua.edu.cn, leliuzhe@163.com,
chenbin2021@hit.edu.cn, taoliangdpg@126.com,
magjhaha@126.com, xiast@sz.tsinghua.edu.cn

Abstract

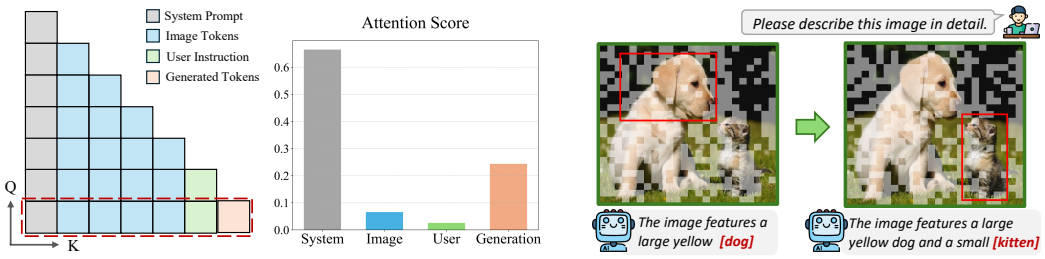
Large Vision-Language Models (LVLMs) are susceptible to hallucinations, where generated responses seem semantically plausible yet exhibit little or no relevance to the input image. Previous studies reveal that this issue primarily stems from LVLMs' over-reliance on language priors while disregarding the visual information during decoding. To alleviate this issue, we introduce a novel Conditional Pointwise Mutual Information (C-PMI) calibrated decoding strategy, which adaptively strengthens the mutual dependency between generated texts and input images to mitigate hallucinations. Unlike existing methods solely focusing on text token sampling, we propose to jointly model the contributions of visual and textual tokens to C-PMI, formulating hallucination mitigation as a bi-level optimization problem aimed at maximizing mutual information. To solve it, we design a token purification mechanism that dynamically regulates the decoding process by sampling text tokens remaining maximally relevant to the given image, while simultaneously refining image tokens most pertinent to the generated response. Extensive experiments across various benchmarks reveal that the proposed method significantly reduces hallucinations in LVLMs while preserving decoding efficiency. The code is available at: https://github.com/ffhibnese/CMI_VLD_Hallucination_Mitigation.

1 Introduction

The unprecedented breakthroughs in large vision-language models (LVLMs) [1, 2, 3, 4, 5] have expanded their applicability across various vision-language (V+L) tasks such as autonomous driving [6, 7]. Benefiting from advanced designs of model architectures and training algorithms, LVLMs trained on high-quality image-text pairs have exhibited outstanding capabilities in cross-modal alignment and complex V+L understanding. Despite the remarkable success, the issue of hallucination continues to pose challenges to LVLMs. Concretely, LVLMs may generate semantically coherent yet factually incorrect contents that are entirely inconsistent with the input image [8, 9, 10]. E.g., describe non-existent objects or misinterpret the attributes and relationships of visual entities within the image. This raises serious concerns regarding the deployment of LVLMs in real-world applications, particularly in high-risk scenarios such as medical diagnosis [11] and financial systems [12].

[*]Equal Contribution
[†]Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Attention allocation. (b) Our visual token refinement.

Figure 1: (a) Illustration of the attention bias of LVLMs. While image tokens constitute the majority of the input tokens, they receive significantly less cumulative attention scores compared to text tokens. (b) The proposed purification mechanism when the masking ratio is 50%. Our method promotes more reliable generation by adaptively retaining image tokens with high relevance to the ongoing response.

To address this issue, one line of research explores fine-tuning for more fine-grained alignment [13, 14] or post-hoc analysis to correct hallucinated elements within the generated responses [15, 16]. Another research stream focuses on directly modifying token distributions during decoding [17, 18, 8, 9, 10, 19]. These methods employ various techniques to penalize the probabilities of hallucination-inducing tokens, thereby encouraging the generation of more faithful and reliable responses. While these decoding-based strategies have shown practical effectiveness, their designs are typically grounded in empirical findings and lack convincing theoretical foundations. Moreover, they generally fail to explicitly quantify and control the dynamic mutual relevance between the visual input and the progressively generated text, leading to insufficient effectiveness in certain scenarios.

In this work, we build upon the line of decoding-based methods and investigate the issue from an information-theoretic perspective, based on which we propose a novel **C**onditional **M**utual **I**nformation-aware adaptive **V**ision-**L**anguage **D**ecoding strategy (CMI-VLD). Specifically, previous studies [20, 21] have revealed that a key factor contributing to hallucination is LVLM's tendency to overly depend on text tokens during the autoregressive generation, with limited attention paid to the critical visual input (see Fig. 1 (a)). As a result, the generated text is guided more by the language priors inherent in the LLM backbone, rather than grounded in the actual visual content of the input image. This eventually leads to a low mutual dependency between the input images and the final responses, hence exacerbating the occurrence of hallucinations in LVLMs.

To mitigate this issue, we introduce conditional pointwise mutual information (C-PMI) to quantify the mutual correlation between the visual inputs and the generated texts during generation. Correspondingly, we reformulate the hallucination mitigation objective as a vision-language mutual information maximization problem, which is further decomposed into two complementary sub-tasks that capture the respective contributions of visual and textual tokens. Based on the analysis, we derive a bi-level optimization formulation and design an effective solution that adaptively calibrates each decoding step during the generation process. To optimize the inner sub-problem, we calibrate the token distribution using the derived formula to prioritize tokens that exhibit strong relevance to the visual input. For the outer sub-problem, we propose an efficient visual token purifier parameterized as a learnable network, to dynamically refine image tokens that are most pertinent to the current textual context. By filtering out redundant image tokens that impair mutual information with the generated content, the proposed strategy directs the model to focus more on the key visual tokens most relevant to the ongoing response (see Fig. 1 (b)), further enhancing the dependence of the generated text on the input image. To summarize, our main contributions are threefold:

- We revisit the hallucination mitigation problem in LVLMs from an information-theoretic perspective, where we reformulate it as a conditional mutual information maximization problem and introduce a novel bi-level optimization-based solution framework.
- To implement this optimization, we propose an effective and efficient adaptive vision-language decoding strategy that dynamically refines the most informative visual and textual tokens to maximize the C-PMI throughout the generation process.
- Extensive experimental results on multiple LVLMs such as LLaVA-1.5 across five evaluation benchmarks demonstrate the exceptional effectiveness of the proposed CMI-VLD in mitigating hallucination, significantly outperforming competitive baselines.

2 Related Work

Large Vision-Language Models. Built upon advanced pre-trained LLMs [22, 23, 24, 25], LVLMs successfully bridge the gap between visual perception and linguistic reasoning [26, 27, 3, 28, 29], achieving impressive performance in generating diverse responses and tackling complex visual understanding tasks. To incorporate visual information into the LLM backbone, LVLMs like LLaVA [2, 4] and Shikra [3] employ linear projection layers trained by instruction fine-tuning to directly map visual features into the LLM embedding space. Meanwhile, the BLIP series [30, 31] introduces Q-former to integrate visual tokens dynamically through gated cross-attention layers, thereby reducing redundancy in image token representations. Benefiting from better training data, improved algorithms, and increasingly powerful LLM backbones, recent LVLMs such as LLaVA-Next [32] have demonstrated stronger multimodal understanding capabilities. Despite the progress, LVLMs still suffer from serious hallucination problems, where the generated responses are plausible yet unfaithful or factually incorrect. Our work aims to mitigate this issue and enhance the reliability of LVLMs.

Mitigating Hallucinations in LVLMs. To address the critical issue, various strategies have been proposed to alleviate hallucinations from different perspectives. Early efforts focused on improving the multimodal alignment by training LVLMs with higher-quality data or more advanced algorithms [13, 14, 33]. However, they often require additional datasets and incur substantial computational overhead, primarily due to the exhaustive instruction-tuning procedures. In parallel, post-hoc correction methods based on auxiliary models have been explored [15, 16] to filter or revise hallucinated content in the output responses. Nevertheless, these methods heavily rely on the performance of the auxiliary model and introduce extra inference overhead.

Another research line focuses on decoding-based hallucination mitigation. These methods primarily seek to construct token distributions that adaptively suppress the probabilities of hallucinated tokens [34, 17, 19, 8, 35]. By sampling from carefully crafted distributions, these methods significantly reduce hallucinated concepts in generated responses. In addition, OPERA [19] identifies a strong correlation between hallucinations and summary tokens, and proposes to penalize the over-trust logits along with a rollback strategy. [20] conducts a modular analysis and empirically reveals that certain attention heads overly focus on textual tokens while neglecting the pivotal visual information, based on which they introduce two correction algorithms to penalize text attentions. Among these methods, only M3ID [21] considers theoretical aspects, yet it introduces mutual information solely to justify its vision-prompt dependency metric in contrastive decoding, without delving deeper into the key factors influencing C-PMI or exploring an effective optimization paradigm. In contrast, this paper proposes a novel multimodal adaptive decoding algorithm grounded in C-PMI, which dynamically amplifies the mutual relevance between image and text and effectively reduces hallucinations in LVLM outputs.

3 Methodology

This section first introduces the basic generative paradigm of LVLMs. Building on this, we propose our adaptive decoding algorithm for hallucination mitigation, *i.e.*, CMI-VLD. Finally, we present the detailed design of a learnable predictor for visual token purification in our method.

3.1 Preliminary

Before delving into the proposed adaptive decoding algorithm, *i.e.*, CMI-VLD, we revisit the autoregressive generation paradigm of LVLMs, which serves as the foundation for subsequent derivations.

Given a user prompt x and an image v as input, a pre-trained LVLM $f_\theta(\cdot)$ first processes the image v through a vision encoder, followed by a cross-modal projection module, to generate a set of visual tokens $v = \{v_0, v_1, \ldots, v_N\}$. At decoding step t, the visual tokens are concatenated with the textual tokens from the instruction x and the previously generated token sequence $y_{<t}$. The resulting sequence is fed into the LLM backbone of the LVLM to autoregressively predict the next token:

$$y_t \sim p_\theta(\cdot \mid v, x, y_{<t}) = \mathrm{softmax}(f_\theta(\cdot \mid v, x, y_{<t})), \tag{1}$$

where y_t is the token being sampled at current generation step t. In particular, the probability of a generated sentence y of length l can be factorized as a product of conditional probabilities:

$$q_\theta(y \mid v, x) = \prod_{t=0}^{l-1} p_\theta(y_t \mid v, x, y_{<t}) = \prod_{t=0}^{l-1} \operatorname{softmax}\left(f_\theta\left(\cdot \mid v, x, y_{<t}\right)\right)_{y_t}, \tag{2}$$

where q_θ denotes the sentence-level conditional probability distribution characterized by the LVLM $f_\theta(\cdot)$. This yields an appealing property for the subsequent expansion of mutual information, as the likelihood of a given text under a specific LVLM can be accurately computed by Eq. (2).

3.2 The Proposed CMI-VLD

To reduce hallucination-related content in the output response, we propose to strengthen the bidirectional dependency between the input image and the generated sentence by maximizing their conditional mutual information measured by the target LVLM $f_\theta(\cdot)$. However, standard CMI computation requires estimating the full conditional distributions of image variable V and text variable Y given the user instruction variable X, which is intractable in practice due to challenges such as dimension explosion or data sparsity. To overcome this challenge, we adopt its pointwise formulation [36], which balances theoretical rigor with practical feasibility, to quantify the local dependency between a specific image–text pair $(V = v, Y = y)$, conditioned on a given instruction $X = x$:

$$\max_{v,y} \text{C-PMI}_\theta(V = v, Y = y \mid X = x) = \max\left(\log \frac{p_\theta(v, y \mid x)}{p_\theta(v \mid x)\, p_\theta(y \mid x)}\right). \tag{3}$$

To achieve more effective optimization, we carefully analyze this objective from the perspectives of both visual and textual data points involved in C-PMI calculation. Given an input image v, the algorithm should encourage the generation of a text y that is highly aligned with the visual input v to strengthen their mutual dependency. Simultaneously, for the given text y, the visual input can be refined to exhibit strong relevance to y, hence further amplifying the mutual information between the two modalities. As a result, the bidirectional dependency between the two variables in maximizing C-PMI naturally induces a bi-level optimization framework, which can be effectively addressed by alternately optimizing the derived inner and outer subproblems. However, the conditional distributions in Eq. (3) can not yet be directly calculated. Based on Eq. (2) and Bayes' Theorem, we then further expand the optimization objective as follows:

$$\max_{v,y} \text{C-PMI}_\theta(v, y \mid x) = \max \sum_{t=0}^{l-1} \left[\log p_\theta\left(y_t \mid v, x, y_{<t}\right) - \log p_\theta\left(y_t \mid x, y_{<t}\right)\right]. \tag{4}$$

Detailed proof is in Appendix A. This formula decomposes the original objective over individual decoding steps, enabling each term to be explicitly computed using the token-level probabilities provided by the LVLM. An interesting observation is that the token distributions used in existing contrastive decoding studies [18, 17, 9, 10] can be viewed as specific variants of the optimization goal in Eq. (4), and thus can be naturally regarded as special cases of our framework when only the text's influence on C-PMI is considered. Next, we concretize the solution of two interdependent subproblems from text and visual modalities to form our alternating optimization procedure:

(1) **Calibrated Distribution Sampling for Text Modality.** To optimize the text sequence y, Eq. (4) encourages us to construct an improved distribution p_c to prioritize text tokens that maximize the difference between probabilities predicted with and without the visual input. However, directly applying this formulation yields unsatisfactory results since it can excessively penalize reasonable tokens in certain contexts. Inspired by [17, 9], we introduce a hyperparameter λ to provide a more fine-grained control over the strength of the subtraction, which can be formally expressed as:

$$y_t \sim p_c(\cdot \mid v, x, y_{<t}) = \operatorname{softmax}\Big[(1+\lambda) f_\theta\left(\cdot \mid v, x, y_{<t}\right) - \lambda f_\theta\left(\cdot \mid x, y_{<t}\right)\Big], \tag{5}$$

This strategy calibrates the token distribution by urging the generation toward tokens that are more informative of the image and hence enhancing its reliance on visual content. To ensure the quality of generated sentences, we also incorporate the adaptive token truncation mechanism [13, 17] to prune the sampling space of Eq. 5 into a more reliable token candidate pool.

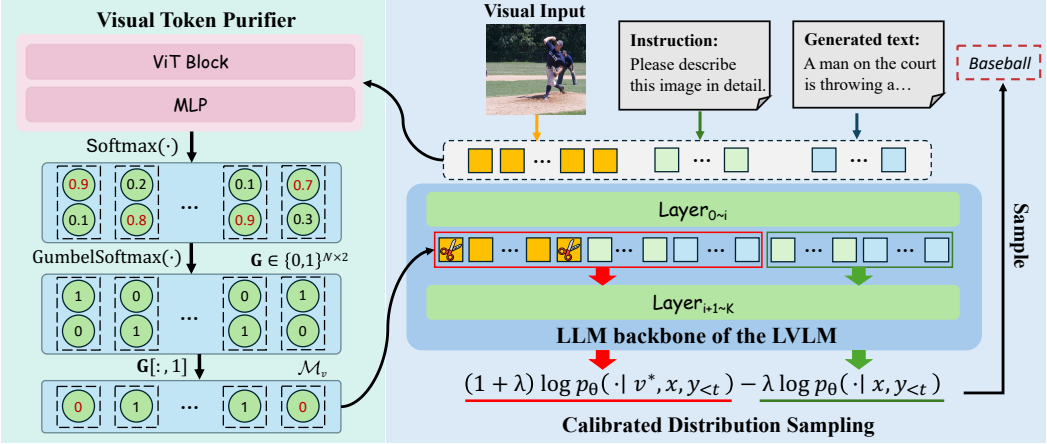


Figure 2: Overview of the proposed CMI-VLD decoding. At each timestep t, CMI-VLD mitigates hallucination by maximizing mutual dependency between the visual input and the ongoing response through the proposed vision-language purification. Specifically, the visual token purifier first incorporates current input tokens to predict an image mask $\mathcal{M}_v$, which filters out irrelevant visual tokens to enhance C-PMI. Based on the refined visual input, a text token distribution is correspondingly constructed to penalize hallucination-related text tokens and hence guide the next-token sampling to further strengthen the dependency on the visual input.

(2) **Visual Token Refinement for Visual Modality.** Motivated by recent findings [37, 38] that many image tokens in LVLMs are redundant, we propose a visual token purification mechanism that enhances C-PMI by evicting tokens considered non-informative with respect to the given text. In this way, the LVLM can focus more on the most critical visual tokens for improved generation. Moreover, we also incorporate the model's attention scores over the visual input to identify tokens that exert a stronger influence on the LVLM's decisions. Given the query vectors $Q_i \in \mathbb{R}^{H\times n\times d_k}$ and key vectors $K_i \in \mathbb{R}^{H\times n\times d_k}$ at the i-th LVLM layer, where H is the head number, n is the current token number, and d_k is the latent dimension, the total attention scores of an image v is calculated as:

$$\text{Attn}_i(v) = \frac{1}{H}\sum_{v_j\in v}\sum_{k=0}^{H-1} A_i^{(k,:,:)}[-1][v_j], \quad \text{where } A_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_k}} + \mathcal{M}_c\right), \tag{6}$$

where $\mathcal{M}_c$ is the causal attention mask and $A_i \in \mathbb{R}^{H\times n\times n}$ denotes the attention matrix at the i-th layer. This design enables the optimizer to select visual tokens that are not only text-relevant but also highly impactful in guiding the model's predictions, boosting the effectiveness of our visual purification. Formally, the overall bi-level optimization objective can be expressed as:

$$\begin{aligned} &\max_{y} \sum_{t=0}^{l-1} \Big[(1+\lambda)\log p_\theta\left(y_t \mid v^*, x, y_{<t}\right) - \lambda \log p_\theta\left(y_t \mid x, y_{<t}\right)\Big], \\ &\text{s.t.} \quad v^* = \arg\max_{v} \Big[\alpha \cdot \text{Attn}_i(v) + \log p_\theta\left(y_t \mid v, x, y_{<t}\right) - \log p_\theta\left(y_t \mid x, y_{<t}\right)\Big], \end{aligned} \tag{7}$$

where λ is the aforementioned correction factor for distribution calibration and α is a non-negative hyperparameter balancing the influence of C-PMI and attention scores on visual token selection.

At decoding step t, we first optimize the upper subproblem by sampling y_t from the distribution adjusted based on Eq. (5). To solve the lower subproblem, we then adaptively retain a proportion γ of image tokens as the purified input to promote its relevance to the current textual context. Motivated by findings in [38] that token sparsification at the second layer of LVLMs yields optimal performance, we utilize attention scores from this layer ($i = 2$) for visual purification and start the refinement accordingly. By alternately solving the two subproblems at each decoding step, the optimizer simultaneously samples text tokens that tightly align with current visual content and purifies informative visual tokens with strong relevance to the ongoing textual context, hence effectively amplifying C-PMI and reducing hallucination-related elements in the final response. For stable performance, we also incorporate the feature steering mechanism [39] into our implementation.

3.3 Visual Token Purifier for Visual Refinement

To address the outer subproblem in Eq. (7), an intuitive solution is to manually select image tokens that maximize the defined score for every decoding step. However, this requires repeatedly calculating token-wise scores and would incur substantial computational burdens compared to existing decoding-based approaches [17, 18]. To overcome this challenge, we propose a lightweight visual token purifier $\mathcal{P}(\cdot)$, which consists of only a few transformer blocks and MLP layers (see Appendix D for details) [37], to automatically filter visual tokens that benefit C-PMI maximization.

As illustrated in Fig. 2, the purifier $\mathcal{P}(\cdot)$ incorporates the concatenated embeddings $\mathbf{z} = [\mathbf{z}_v, \mathbf{z}_x, \mathbf{z}_{y_{<t}}]$ of the image v and the current text $(x, y_{<t})$ to output a probability distribution $\boldsymbol{\pi} = \mathrm{softmax}\left(\mathcal{P}\left(\mathbf{z}\right)\right) \in [0,1]^{N\times 2}$, where N is the number of visual tokens. Here, $\boldsymbol{\pi}_{i,0}$ represents the probability of discarding the i-th token, and $\boldsymbol{\pi}_{i,1}$ represents the probability of retaining it. The final visual token mask $\mathcal{M}_v \in \{0,1\}^N$ can be then extracted via:

$$\mathcal{M}_v = \left\{ \underset{j\in\{0,1\}}{\arg\max}\ \boldsymbol{\pi}_{ij} \;\middle|\; i \in \{0,1,\ldots,N-1\} \right\}. \tag{8}$$

Model Training. The principle challenge in training the purifier lies in the non-differentiability of the $\arg\max$ operation used for discrete token selection. To address this, we employ the Gumbel-Softmax technique with a temperature parameter τ to enable differentiable sampling:

$$\mathbf{G} = \text{Gumbel-Softmax}(\boldsymbol{\pi}, \tau),$$

where the sampling output $\mathbf{G} \in \{0,1\}^{N\times 2}$ containing N one-hot vectors. Since the retention probability of a visual token corresponds to the second column in $\boldsymbol{\pi}$, the differentiable mask $\mathcal{M}_v \in \{0,1\}^N$ can be extracted as $\mathcal{M}_v = \mathbf{G}[:,1]$. This approach introduces stochasticity via noise from a fixed Gumbel distribution, which enables gradients to propagate back through the probability parameters. Moreover, the temperature factor τ helps soften the sampling distribution, thereby improving gradient stability and facilitating convergence during training.

To specify the retaining ratio γ, we introduce a Frobenius norm-based regularization term that penalizes incorrect retention rate. The overall training objective at decoding step t is defined as:

$$\begin{aligned}\mathcal{L}_{total} = \big(\log p_\theta(y_t \mid v, x, y_{<t}) - \log p_\theta(y_t \mid x, y_{<t})\big) \\ +\alpha\cdot\text{Attn}_i(v) + \beta\cdot\left\|\mathrm{sum}(M_v)/N - \gamma\right\|_F,\end{aligned} \tag{9}$$

where β is a weight coefficient controlling the regularization strength of the reduction ratio, $\|\cdot\|_F$ denotes the F-norm of a matrix, and $\mathrm{sum}(\cdot)$ represents the summation operation.

By iteratively updating the network using the loss function in Eq. (9) on paired image-text data, the purifier learns to dynamically identify visual tokens that effectively contribute to mutual information maximization, which further enhances the informativeness of the visual input while discarding those redundant and distracting visual tokens. Furthermore, our method preserves the decoding efficiency despite introducing an additional network, as the purifier module is lightweight and the removal of non-essential visual tokens helps reduce the overall inference cost (see Sec . 4.2).

4 Experiments

4.1 Experimental Setup

Models and Baselines. We align with [9] and choose four representative LVLMs for evaluation, including InstructBLIP [31], Shikra [3], LLaVA-1.5 on the 7B scale [4], and LLaVA-NeXT [32] on the 8B scale. We conduct a comprehensive evaluation of the proposed CMI-VLD on a range of state-of-the-art (SOTA) baselines, including Sampling (Top-p=1), Greedy, VTI [39], VCD [17], ICD [18], HALC [8], OPERA [19], SID [9], and VASparse [10]. Following SID, we implement the proposed method under both sampling and greedy decoding settings. It is worth noting that HALC, OPERA, and VASparse adopt the more flexible and stronger beam search strategy, which may raise fairness concerns as it retains a broader set of promising candidate paths during decoding. Nevertheless, our CMI-VLD still consistently outperforms these methods.

Table 1: Comparison of the proposed CMI-VLD with SOTA baselines on the CHAIR metric. We evaluate the performance on MSCOCO. The † indicates decoding strategies based on beam search.

Method	LLaVA-1.5		InstructBLIP		Shikra		LLaVA-Next	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Sampling	52.2	15.8	55.0	25.3	56.2	15.8	34.8	9.4
ICD	51.0	15.2	64.0	20.2	56.6	15.5	33.4	8.7
VCD	50.4	15.6	57.6	19.2	56.2	15.5	36.0	9.3
VTI	37.2	11.4	**49.2**	21.9	47.0	14.1	32.2	7.8
SID	49.2	14.5	58.0	18.7	54.4	14.4	39.4	9.9
CMI-VLD	**30.2**	**9.3**	51.0	**16.1**	**38.2**	**10.1**	**30.6**	**7.6**
Greedy	45.0	13.5	52.2	21.8	54.8	15.8	31.6	8.2
ICD	44.8	12.8	48.8	14.1	55.0	14.0	32.8	9.1
VCD	49.4	14.0	46.6	13.3	55.8	15.3	36.8	9.4
HALC†	33.2	10.3	61.4	20.0	55.4	14.7	36.7	9.5
OPERA†	39.4	10.3	48.2	13.8	36.8	11.7	33.6	8.3
VTI	30.6	10.1	48.3	20.7	44.6	13.7	30.1	7.6
SID	42.8	12.1	56.2	15.8	51.2	13.6	38.0	8.9
VASparse†	49.6	14.2	53.6	14.9	51.6	14.8	33.6	9.1
CMI-VLD	**29.9**	**8.9**	**43.2**	**12.9**	**30.6**	**8.9**	**27.2**	**6.8**

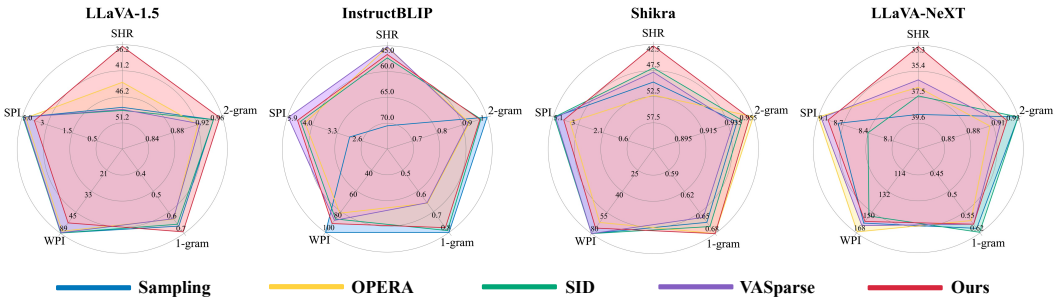


Figure 3: GPT-4o assisted benchmark. We calculate the Sentence-level Hallucination Ratio (SHR) as the major metric for hallucination degree, along with 1&2-gram, the number of sentences per image (SPI), and the number of words per image (WPI). A larger radar area indicates better performance.

Evaluation Benchmarks. Following the evaluation protocol in [9, 35], we analyze our CMI-VLD across five widely-used benchmarks: (1) the CHAIR metric [40] on the MSCOCO dataset [41] that measures object hallucinations; (2) a GPT-4 assisted evaluation [33], where we adopt the advanced GPT-4o [42] to detect more fine-grained hallucinations and compute the Sentence-level Hallucination Ratio (SHR); (3) Polling-based Object Probing Evaluation (POPE) [43], another object hallucination evaluation also conducted on MSCOCO; (4) Multimodal Large Language Model Evaluation (MME) [44], a general-purpose benchmark for assessing multimodal capabilities; and (5) MMBench [45], which includes multiple-choice questions designed to evaluate visual perception and reasoning.

Implementation Details. We set $\alpha = 1 \times 10^2$ and $\beta = 5 \times 10^2$ in Eq. (9) during the training of the purifier for all LVLMs. To initiate the refinement process, we adopt Layer $i = 2$ for LLaVA-1.5, Shikra, and LLaVA-NeXT and $i = 4$ for InstructBLIP. In Sec. 4.3, we explore the contrastive strength λ ranging from 0 to 0.9. For all experiments, we set *max new tokens* as 512 for evaluation. Note that CMI-VLD is compatible with the feature steering mechanism in VTI [39], which is then incorporated in our implementation for stable and enhanced performance. More details are in Appendix B.

4.2 Performance Evaluation

CHAIR Evaluation. Following previous studies [19, 9, 8], we query LVLMs with the input prompt "`Please describe this image in detail.`" using 500 images randomly sampled from the validation set of MSCOCO. By dynamically amplifying the mutual relevance between visual inputs and generated texts, the proposed method achieves remarkable improvements over SOTA baselines

Table 2: Comparison of the proposed CMI-VLD with SOTA baselines on the POPE metric. The † indicates decoding strategies based on beam search.

Model	Method	Random		Popular		Adversarial	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
LLaVA-Next	*Sampling*	82.53%	79.19%	81.57%	78.31%	80.30%	77.16%
	ICD	82.77%	79.57%	81.97%	78.81%	81.03%	77.95%
	VCD	83.67%	80.80%	82.17%	79.40%	80.90%	78.25%
	VTI	82.70%	79.45%	81.43%	78.25%	80.10%	77.05%
	SID	84.67%	82.20%	83.57%	81.16%	81.60%	80.27%
	CMI-VLD	**85.17%**	**83.02%**	**84.10%**	**82.03%**	**82.30%**	**80.40%**
	Greedy	83.40%	80.32%	82.60%	79.55%	81.77%	78.77%
	ICD	83.47%	80.41%	82.60%	79.56%	81.90%	78.91%
	VCD	84.43%	81.85%	83.30%	80.77%	82.33%	79.88%
	HALC†	83.34%	80.36%	82.33%	79.48%	81.40%	78.92 %
	OPERA†	83.50%	80.46%	82.70%	79.69%	81.87%	78.91%
	VTI	84.70%	82.09%	83.67%	81.11%	82.90%	80.40%
	SID	84.97%	82.53%	83.93%	81.56%	82.97%	80.67%
	VASparse†	83.47%	80.52 %	82.24 %	79.69%	81.33%	78.88 %
	CMI-VLD	**86.43%**	**84.52%**	**85.07%**	**83.22%**	**83.90%**	**82.14%**
InstructBLIP	*Sampling*	82.03%	81.30%	78.77%	78.66%	76.37%	76.81%
	VTI	83.50%	82.01%	80.83%	79.70%	79.13%	78.29%
	ICD	83.20%	82.29%	79.87%	79.51%	77.63%	77.74%
	VCD	83.43%	82.49%	79.70%	79.36%	77.53%	77.65%
	SID	85.43%	84.81%	82.43%	82.24%	79.47%	79.84%
	CMI-VLD	**86.33%**	**85.41%**	**84.60%**	**83.87%**	**81.57%**	**81.29%**
	Greedy	87.27%	85.91%	84.87%	83.72%	82.97%	82.04%
	ICD	87.23%	85.82%	84.90%	83.68%	**83.13%**	82.11%
	VCD	86.73%	85.30%	84.37%	83.16%	82.47%	81.49%
	HALC†	87.30%	85.96%	84.83%	83.70%	83.00%	82.08%
	OPERA†	87.53%	86.26%	85.07%	84.00%	83.07%	82.24%
	VTI	85.73%	83.86%	84.13%	82.36%	82.50%	80.89%
	SID	88.10%	87.15%	85.87%	85.10%	82.90%	82.52%
	VASparse†	87.33%	86.00%	84.87%	83.74%	83.00%	82.09%
	CMI-VLD	**88.37%**	**87.50%**	**86.10%**	**85.40%**	82.87%	**82.64%**

on different LVLMs, as observed in Table 1. *E.g.*, a notable improvement of 7% and 2.1% in C_S and C_I for *Sampling* on the LLaVA-1.5 model. Notably, some baselines even exacerbate hallucination content compared to standard decoding strategies in some cases. In contrast, the proposed CMI-VLD consistently reduces both sentence-level and instance-level object hallucinations in the final responses.

GPT-4o Assisted Evaluation. While CHAIR is a reliable evaluation metric widely adopted in previous studies, it is limited within the scope of object hallucinations and fails to identify other types, such as attribute, relational, and positional hallucinations. To more comprehensively evaluate the effectiveness of our method, we introduce the GPT-assisted benchmark [33], which uses the object-level descriptions in the Visual Genome dataset [46] as ground-truth, to judge more fine-grained hallucinations assisted by the advanced GPT-4o. Figure 3 demonstrates that the proposed CMI-VLD significantly outperforms SOTA baselines across four LVLMs. Compared with competitive baselines, we achieve a relative improvement of 15.89% for LLaVA-1.5 in the hallucination metric SHR while maintaining text fluency. We also note that our method reduces the length of generated texts to some extent, which can be caused by the removal of hallucinated sentences [19].

POPE Evaluation. The POPE metric also focuses on object hallucinations by using a prompt "`Is there a <object> in the image?`" to query LVLMs for answering a yes/no question. We report the results of the accuracy and F1 score in Table 2. The quantitative results reveal that our method generally performs best across the three split datasets. Notably, in the POPE evaluation, where responses are typically short and follow fixed patterns such as "`Yes, there is a <object> in the image.`", the evaluation primarily hinges on the first one or few tokens (*i.e.*, Yes or No) [19].

Table 3: Comparison of the proposed CMI-VLD with SOTA baselines on LVLM benchmarks. The † indicates beam search-based methods, while other methods adopt the same *Greedy* decoding.

Benchmarks	*Greedy*	ICD	VCD	HALC†	OPERA†	VTI	SID	VASparse†	CMI-VLD
MME	1465.11	1432.43	1472.57	1473.43	1471.37	1435.35	1467.05	1466.61	**1481.17**
MMBench	64.86	64.26	64.26	57.90	64.78	64.43	64.26	64.78	**65.12**

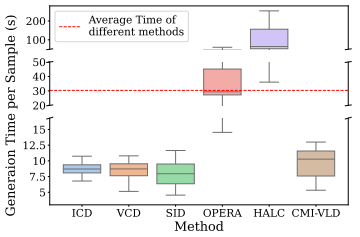


Figure 4: Generation time per sample of different methods.

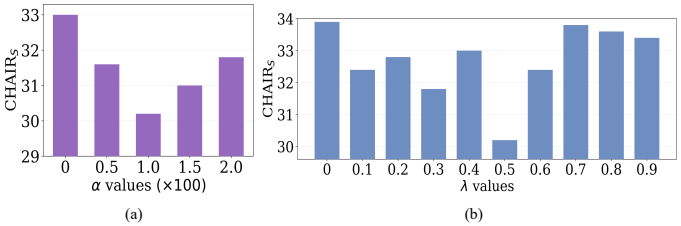


Figure 5: CHAIR$_S$ results of the proposed CMI-VLD under varying values of hyperparameters α and λ.

Consequently, our CMI-VLD may not fully exhibit its effectiveness in this constrained setup, as it is designed to dynamically adjust decoding over the entire generation rather than concentrating solely on the initial tokens. Nevertheless, our method still achieves notable improvements over competitive baselines. Due to the page limit, results on more LVLMs are provided in Appendix C.

MME and MMBench Evaluations. Apart from the above benchmarks tailored for hallucination evaluation, we additionally test on two popular LVLM benchmarks, *i.e.*, MME [7] and MMBench [45], to systematically analyze their various capability dimensions. MME provides a suite of fine-grained, image-grounded multiple-choice questions across various categories. We follow SID and report the overall perception score covering 10 sub-tasks, such as object existence and fine-grained recognition. MMBench is another large-scale bilingual benchmark consisting of over 3,000 curated multiple-choice questions. We compute the LVLM's average score across 20 multimodal tasks, such as attributes, logical reasoning, and coarse/fine-grained perception, to comprehensively evaluate its capabilities. As observed, CMI-VLD not only reduces the hallucinated contents but also enhances diverse capabilities of MLLMs, bringing remarkable improvements over the default decoding methods. These results underscore CMI-VLD as a reliable and practical strategy for hallucination mitigation.

Inference Time Analysis. Since our method introduces an additional visual token purifier for effective visual refinement, it is crucial to assess its influence on the overall computational efficiency. Following [20], we calculate the generation time per response based on LLaVA-1.5 to assess computing burdens. Figure 4 reveals that the proposed CMI-VLD achieves satisfactory decoding efficiency, introducing negligible computational overhead. This is primarily attributed to the lightweight architecture of the visual purifier and the removal of redundant visual tokens that would incur significant computational overhead, demonstrating that CMI-VLD effectively balances performance and efficiency.

4.3 Ablation Study

Next, we provide ablation analysis regarding several critical hyperparameters. More ablation analysis about the retaining ratio and the effectiveness of the proposed techniques is presented in Appendix C.

The effect of varying loss parameter α. The value of α is a critical factor as it adjusts the contribution of the attention scores during purifier training. We then evaluate the performance under various values of α in Figure 5 (a). The performance gains observed when comparing to $\alpha = 0$ suggest that incorporating $\text{Attn}_i(\cdot)$ enhances the effectiveness of the visual purifier. Moreover, the results indicate that $\alpha = 1 \times 10^2$ yields the best performance, and is therefore adopted in our main experiments.

The effect of calibration intensity λ. During decoding, the hyperparameter λ plays a pivotal role in regulating the strength of distribution calibration. We present the CHAIR results under varying λ in Figure 5 (b). The proposed method reaches optimal performance when $\lambda = 0.5$, hence we adopt it as the default setup. Moreover, we emphasize that a properly selected range of positive values of λ yields significant improvements over the $\lambda = 0$ setup, validating our distribution calibration strategy.

5 Conclusion

In this work, we first revisit the key reason for hallucination in LVLMs, based on which we introduce conditional mutual information as a theoretical foundation to enhance the mutual dependency between visual input and generated text. To strengthen this cross-modal association, we propose a novel CMI-aware bi-level optimization framework, which is efficiently and effectively solved via a carefully designed vision-language decoding strategy. Through extensive experiments across multiple LVLMs and evaluation benchmarks, we demonstrate the superiority of the proposed approach in mitigating hallucinations and improving the recognition capability of LVLMs in diverse scenarios.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under grant 62171248, 62301189, 62576122, and Shenzhen Science and Technology Program under Grant KJZD20240903103702004, JCYJ20220818101012025, GXWD20220811172936001.

References

[1] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," in *The Twelfth International Conference on Learning Representations*.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

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

[4] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[5] H. Fang, J. Kong, W. Yu, B. Chen, J. Li, H. Wu, S. Xia, and K. Xu, "One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models," *arXiv preprint arXiv:2406.05491*, 2024.

[6] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[7] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 958–979.

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

[9] F. Huo, W. Xu, Z. Zhang, H. Wang, Z. Chen, and P. Zhao, "Self-introspective decoding: Alleviating hallucinations for large vision-language models," *arXiv preprint arXiv:2408.02032*, 2024.

[10] X. Zhuang, Z. Zhu, Y. Xie, L. Liang, and Y. Zou, "Vasparse: Towards efficient visual hallucination mitigation for large vision-language model via visual-aware sparsification," *arXiv preprint arXiv:2501.06553*, 2025.

[11] F. Liu, T. Zhu, X. Wu, B. Yang, C. You, C. Wang, L. Lu, Z. Liu, Y. Zheng, X. Sun *et al.*, "A medical multimodal large language model for future pandemics," *NPJ Digital Medicine*, vol. 6, no. 1, p. 226, 2023.

[12] J. Huang, M. Xiao, D. Li, Z. Jiang, Y. Yang, Y. Zhang, L. Qian, Y. Wang, X. Peng, Y. Ren *et al.*, "Open-finllms: Open multimodal large language models for financial applications," *arXiv preprint arXiv:2408.11878*, 2024.

[13] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating hallucination in large multi-modal models via robust instruction tuning," *arXiv preprint arXiv:2306.14565*, 2023.

[14] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun *et al.*, "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 807–13 816.

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

[16] S. Yin, C. Fu, S. Zhao, T. Xu, H. Wang, D. Sui, Y. Shen, K. Li, X. Sun, and E. Chen, "Woodpecker: Hallucination correction for multimodal large language models," *Science China Information Sciences*, vol. 67, no. 12, p. 220105, 2024.

[17] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing, "Mitigating object hallucinations in large vision-language models through visual contrastive decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 872–13 882.

[18] X. Wang, J. Pan, L. Ding, and C. Biemann, "Mitigating hallucinations in large vision-language models with instruction contrastive decoding," in *ACL (Findings)*, 2024.

[19] Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu, "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 418–13 427.

[20] T. Yang, Z. Li, J. Cao, and C. Xu, "Mitigating hallucination in large vision-language models via modular attribution and intervention," in *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

[21] A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto, "Multi-modal hallucination control by visual information grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 303–14 312.

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

[23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[24] J. Kong, H. Fang, X. Yang, K. Gao, B. Chen, S.-T. Xia, Y. Wang, and M. Zhang, "Wolf hidden in sheep's conversations: Toward harmless data-based backdoor attacks for jailbreaking large language models," *arXiv preprint arXiv:2505.17601*, 2025.

[25] H. Fang, J. Kong, T. Zhuang, Y. Qiu, K. Gao, B. Chen, S.-T. Xia, Y. Wang, and M. Zhang, "Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors," *arXiv preprint arXiv:2505.15337*, 2025.

[26] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi *et al.*, "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.

[27] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[28] H. Fang, J. Kong, B. Chen, T. Dai, H. Wu, and S.-T. Xia, "Clip-guided generative networks for transferable targeted adversarial attacks," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–19.

[29] T. Zhang, K. Gao, J. Bai, L. Y. Zhang, X. Yin, Z. Wang, S. Ji, and W. Chen, "Pre-training clip against data poisoning with optimal transport-based matching and alignment," *arXiv preprint arXiv:2509.18717*, 2025.

[30] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[31] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2305.06500

[32] B. Li and F. Huo, "Reqa: Coarse-to-fine assessment of image quality to alleviate the range effect," *Journal of Visual Communication and Image Representation*, vol. 98, p. 104043, 2024.

[33] Z. Zhao, B. Wang, L. Ouyang, X. Dong, J. Wang, and C. He, "Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization," *arXiv preprint arXiv:2311.16839*, 2023.

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

[35] C. Wang, X. Chen, N. Zhang, B. Tian, H. Xu, S. Deng, and H. Chen, "Mllm can see? dynamic correction decoding for hallucination mitigation," *International Conference on Learning Representations*, 2025.

[36] L. Van Der Poel, R. Cotterell, and C. Meister, "Mutual information alleviates hallucinations in abstractive summarization," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 5956–5965.

[37] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, pp. 13 937–13 949, 2021.

[38] L. Chen, H. Zhao, T. Liu, S. Bai, J. Lin, C. Zhou, and B. Chang, "An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 19–35.

[39] S. Liu, H. Ye, and J. Zou, "Reducing hallucinations in large vision-language models via latent space steering," in *The Thirteenth International Conference on Learning Representations*, 2025.

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

[41] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 2014, pp. 740–755.

[42] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

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

[44] Z. Liang, Y. Xu, Y. Hong, P. Shang, Q. Wang, Q. Fu, and K. Liu, "A survey of multimodel large language models," in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 2024, pp. 405–409.

[45] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European conference on computer vision*. Springer, 2024, pp. 216–233.

[46] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[47] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–387.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our contributions and research scope in the claims presented in the abstract and introduction, all of which are supported by theoretical foundation or extensive experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We present the limitations of our paper in a section of the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided full proof for the formulas in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed information of experiments are provided in the Experimental section and the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit our code in the supplementary material and will open-source the code, models, and data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide exhaustive training and test details in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide hyperparameter analysis in different settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide them in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform every respect of the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We highlight that our work can reduce hallucinations in LVLMs and enhance their reliability in real-world applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The license and terms of use are properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets introduced except code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We do not utilize LLMs for our core method development.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Complete Derivation of Equation (4)

Based on Eq. (2) and Bayes' theorem, we provide the detailed derivation of Eq. (4) as follows:

$$\max_{v,y} \text{C-PMI}_\theta(v, y \mid x) = \max_{v,y} \log \frac{p_\theta(v, y \mid x)}{p_\theta(v \mid x)\, p_\theta(y \mid x)} \tag{10}$$

$$= \max_{v,y} \log \frac{p_\theta(v, x, y)/p_\theta(x)}{p_\theta(v \mid x)\, p_\theta(y \mid x)} \tag{11}$$

$$= \max_{v,y} \log \frac{p_\theta(v, x, y)}{p_\theta(x)\, p_\theta(v \mid x)\, p_\theta(y \mid x)} \tag{12}$$

$$= \max_{v,y} \log \frac{p_\theta(v, x, y)}{p_\theta(v, x)\, p_\theta(y \mid x)} \tag{13}$$

$$= \max_{v,y} \log \frac{p_\theta(y \mid v, x)}{p_\theta(y \mid x)} \tag{14}$$

$$= \max_{v,y} \log \frac{\prod_{t=0}^{l-1} p_\theta(y_t \mid v, x, y_{<t})}{\prod_{t=0}^{l-1} p_\theta(y_t \mid x, y_{<t}))} \tag{15}$$

$$= \max_{v,y} \log \prod_{t=0}^{l-1} p_\theta(y_t \mid v, x, y_{<t}) - \log \prod_{t=0}^{l-1} p_\theta(y_t \mid x, y_{<t})) \tag{16}$$

$$= \max_{v,y} \sum_{t=0}^{l-1} \left[\log p_\theta(y_t \mid v, x, y_{<t}) - \log p_\theta(y_t \mid x, y_{<t})\right]. \tag{17}$$

B Experimental Details

B.1 Implementation Details

Throughout our experiments, we retain 80% of the visual input for LLaVA and LLaVA-NeXT, and 90% for Shikra and InstructBLIP. To guide the training of the purifier, we utilize image-text pairs from ShareGPT4V [47]—a high-quality image question answering dataset constructed using images from the MSCOCO dataset. Specifically, we use 2,000 samples for training the purifiers of LLaVA and LLaVA-NeXT, and 4,000 samples for InstructBLIP and Shikra. The learning rate is set to 1×10^{-6} across all models for the decoding hyperparameters of LLMs, and the purifier is trained for 5 epochs.

B.2 Evaluation Model

As mentioned above, we adopt InstructBLIP [31], Shikra [3], LLaVA-1.5 on the 7B scale [4], and LLaVA-NeXT [32] on the 8B scale. InstructBLIP employs Q-former as a cross-modal connector, leveraging 32 learned query tokens to extract and align visual features with text representations in an efficient manner. Other models adopt a simpler architecture of linear projection layers, which directly map visual features into the language model's embedding space, typically using a larger number of image tokens (256 or even 576) as input.

B.3 Evaluation Benchmarks

CHAIR Evaluations. The Caption Hallucination Assessment with Image Relevance (CHAIR) metric is specifically designed to evaluate object hallucination in image captioning tasks. It quantifies the extent to which a generated caption includes references to objects that are not present in the corresponding ground-truth annotations. Specifically, CHAIR computes the proportion of hallucinated objects, those mentioned in the generated caption but absent from the reference object set, providing a direct measure of hallucination severity. CHAIR comprises two commonly used variants: CHAIR_i (C_I) and CHAIR_s (C_S), which evaluate the degree of object hallucination at the instance and sentence level, respectively. The lower values of C_I and C_S correspond to a lower degree of object hallucination, indicating greater factual consistency. The two variants can be formulated as follows:

$$C_I = \frac{|\text{hallucinated objects}|}{|\text{all mentioned objects}|}, \quad C_S = \frac{|\text{captions with hallucinated objects}|}{|\text{all captions}|}$$

POPE Evaluations. The Polling-based Object Probing Evaluation (POPE) benchmark is also proposed to evaluate object hallucination in LVLMs. It adopts a discriminative approach by prompting models with binary questions such as "Is there a <object> in the image?" to assess whether the model can correctly identify the presence or absence of specific objects. To ensure balanced evaluation, POPE includes a 50%/50% ratio of queries about present and absent objects. POPE further categorizes the queries into three negative sampling settings: (1) *random*, where absent objects are sampled randomly; (2) *popular*, where negative objects are selected from the most frequent categories; (3) *adversarial*, where negative objects are chosen based on their high co-occurrence likelihood with present ones to increase difficulty. Evaluation is conducted using Accuracy and F1 score, with higher scores indicating stronger performance in mitigating object hallucinations. Due to the concise format of POPE responses, which are typically short declarative sentences, the benchmark primarily reflects the visual grounding ability of a model rather than its long-form generation capacity.

GPT-4 Assisted Evaluations. In addition to object-level hallucinations via CHAIR and POPE, we adopt the GPT-4 assisted benchmark [33], which leverages fine-grained object-level annotations from the Visual Genome (VG) dataset [46] as ground truth. In our implementation, we employ the advanced GPT-4o to identify detailed hallucinations, such as positional, relational, and attribute-based errors, and compute the Sentence-level Hallucination Ratio (SHR) as evaluation results. Given the generated captions and manually annotated facts, GPT-4o is prompted by a template to assess hallucinations for every sentence. Following previous studies [19, 9], we evaluate on 200 VG images with a maximum output length of 512 tokens based on the prompt: "`Please describe this image in detail.`"

MME and MMBench Evaluations. MLLM Evaluation (MME) benchmark is designed to rigorously assess hallucination in MLLMs. It provides a suite of fine-grained, image-grounded multiple-choice questions across various categories, such as object recognition, OCR, counting, and commonsense reasoning, each requiring accurate visual understanding. By offering carefully controlled distractors and a consistent answer format, MME allows for precise evaluation of a model's ability to generate faithful, image-grounded responses. MMBench is a large-scale, bilingual, multimodal benchmark designed to comprehensively evaluate the capabilities of vision-language models (VLMs). It consists of over 3,000 carefully curated multiple-choice questions covering 20 fine-grained ability dimensions, ranging from perception to reasoning. To ensure robustness and fairness, MMBench introduces the CircularEval strategy, where models must consistently answer a question across multiple permutations of choices. MME and MMBench provide rigorous and scalable frameworks for evaluating multimodal understanding and instruction-following capabilities across a wide spectrum of models.

B.4 Principles of Hyperparameter Choices and Adaptation to New LVLMs.

The hyperparameters are chosen based on our empirical analysis and relevant literature, which are further supported through ablation studies. Specifically, we summarize our choice strategy behind several critical hyperparameters for LLaVA-1.5 in our algorithm as follows:

- Contrast strength λ balances the difference between the vision-conditioned and vision-free distributions. Large values may favor casual and incorrect tokens, while a small λ can fail to sufficiently penalize hallucination-prone tokens. We aim to preserve correct distributions while penalizing hallucinated predictions, and thus select a moderate value of $\lambda = 0.5$, which is further validated by ablations (see Fig. 5 (b)) and prior studies [9, 17].
- Visual token retention ratio γ is a sensitive and critical parameter. A high γ may retain noisy tokens and weaken the C-PMI maximization, while a low can discard important visual information. Hence, we adopt an adaptive strategy: for models with many visual tokens (e.g., LLaVA-1.5), we set a relatively smaller $\gamma = 80\%$; for models with fewer, already refined tokens (e.g., InstructBLIP), we use a higher $\gamma = 90\%$. Ablation studies on each model validate the rationality, with results on LLaVA-1.5 shown in Fig. 7 as an illustration.
- Coefficient of attention loss α balances the importance between attention loss and C-PMI loss during purifier training. Empirically, we find the attention loss to be $\sim$1000x smaller in magnitude compared to C-PMI loss, so we set $\alpha = 100$ to adequately amplify its impact while preserving the dominance of the C-PMI objective.
- Purification starting layer i is chosen based on existing well-established studies on token selection [9, 38], which have been empirically validated to yield strong task-specific performance while preserving its general capability.

Table 4: Comparison of the proposed CMI-VLD with SOTA baselines on the POPE metric. To make a fair comparison, all the methods are based on the *sampling* decoding.

Model	Method	Random		Popular		Adversarial	
		Accuracy	**F1 score**	**Accuracy**	**F1 score**	**Accuracy**	**F1 score**
LLaVA-1.5	Default	85.20%	85.42%	81.67%	82.50%	76.20%	78.40%
	ICD	85.73%	85.84%	81.90%	82.61%	76.70%	78.68%
	VCD	83.77%	84.24%	80.77%	81.84%	76.10%	78.38%
	VTI	85.23%	85.33%	82.77%	83.24%	76.63%	78.56%
	SID	87.93%	87.65%	84.57%	84.69%	79.43%	80.59%
	Ours	**88.63%**	**87.83%**	**86.37%**	**85.71%**	**82.27%**	**82.18%**
Shikra	Default	85.07%	83.44%	83.13%	81.68%	81.63%	80.37%
	ICD	85.27%	83.87%	83.13%	81.94%	81.73%	80.73%
	VCD	85.17%	83.77%	83.27%	82.03%	**82.03%**	80.96%
	VTI	84.03%	81.94%	82.43%	80.49%	81.07%	79.29%
	SID	85.53%	84.26%	83.47%	82.39%	81.43%	80.64%
	Ours	**86.23%**	**85.15%**	**83.83%**	**82.96%**	81.63%	**81.08%**

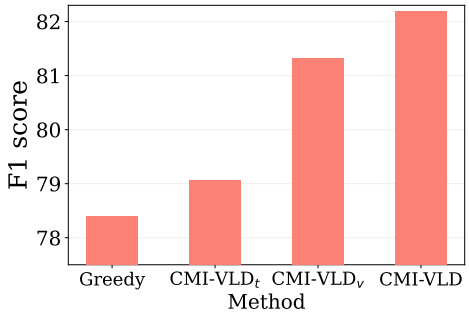


Figure 6: Ablation analysis of the proposed two techniques on the POPE metric.

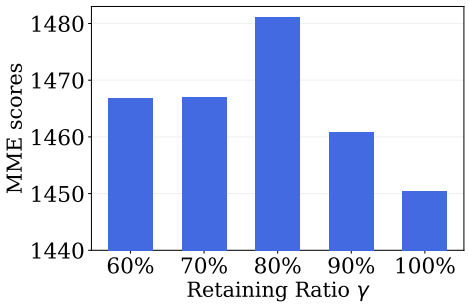


Figure 7: MME results of the proposed CMI-VLD under varying values of retaining ratio γ.

For the adaptation of our method, we empirically observe that the above hyperparameters can transfer to new LVLMs and achieve effectiveness. A necessary adaptation involves slightly tuning λ according to the characteristics of the new models. We also recommend adjusting the retention ratio γ based on the number of input visual tokens, as previously suggested. In addition, it is also necessary to correspondingly adjust the hyperparameters in latent feature steering [39] for different LVLMs.

B.5 Explanation about the Inference Time Comparison

Notably, following SID [9], we evict the selected visual tokens by applying masks to the attention matrices for implementation convenience, rather than physically removing them. The inference time of CMI-VLD reported in Fig. 4 is measured under this implementation and demonstrates that our method maintains inference efficiency. In practice, performing actual eviction of visual tokens would further accelerate inference, implying that the real efficiency advantage of CMI-VLD is even greater.

C More Experimental Results

POPE evaluation on more LVLMs. We supplement the results of POPE metrics on more LVLMs, including LLaVA-1.5 and Shikra. The quantitative results in Table 4 again confirm the effectiveness of our method in mitigating object hallucinations.

Ablation study of vision-language decoding. We then conduct an ablation analysis to validate the contributions of the proposed two techniques, *i.e.*, *Calibrated Distribution Sampling* and *Visual Token Refinement*, which interact with each other to fully maximize the C-PMI. Specifically, we design two variants CMI-VLD$_t$ and CMI-VLD$_v$, which retain only the *Calibrated Distribution Sampling*

Table 5: Inference costs under varying sequence lengths. The number of visual tokens is fixed to 576.

Metric	Method	Sequence Length				
		633 (prefilling)	850	1000	1250	1500
FLOPs (1e14)	w/o purifier	1.9214	2.5814	3.0391	3.8044	4.5731
	CMI-VLD	1.5671	2.4582	3.1226	4.3182	5.6241
Inference Latency (s)	w/o purifier	0.37	17.99	30.26	50.43	70.72
	CMI-VLD	0.35	18.53	31.17	52.2	73.22

Table 6: Inference costs under varying numbers of input visual tokens. We use a fixed text query from the CHAIR evaluation, where the number of text tokens is 56.

Metric	Method	Visual Token Count			
		49	256	576	1024
FLOPs (1e14)	w/o purifier	0.3188	0.9448	1.9214	3.3067
	CMI-VLD	0.2888	0.7876	1.5671	2.6718
Inference Latency (s)	w/o purifier	0.24	0.28	0.37	0.60
	CMI-VLD	0.23	0.28	0.35	0.53

and *Visual Token Refinement*, respectively. Results in Fig. 6 reveal that both the removal of the two components degrade the performance of our algorithm, validating their considerable contributions to guarantee a successful approach for hallucination mitigation.

Ablation study of varying retaining ratio. The retaining ratio γ is a sensitive hyperparameter that should be carefully tuned. A high retaining ratio may fail to sufficiently enhance C-PMI, whereas an excessively low value can degrade performance due to information loss. We evaluate the influence under varying γ to confirm the optimal value. Fig. 7 indicates that $\gamma = 80\%$ is an optimal choice.

Detailed Analysis of the Computational Costs. To reduce computational overhead, we design the purifier as a lightweight network with only 0.1% of the parameters of the LVLM. Its effectiveness in mitigating hallucination has been thoroughly validated by extensive experiments in the main text. Next, we present a detailed computational cost analysis of the visual purifier using LLaVA-Next 8B.

As shown in Table 5 and 6, thanks to its lightweight design and visual token reduction, our purifier introduces negligible overhead and generally maintains computational efficiency comparable to the purifier-free variant. Furthermore, as the number of visual tokens increases, the benefits of visual token reduction become more pronounced, further reducing the computational complexity.

Table 7: CHAIR metrics and Token-per-second (TPS) of CMI-VLD and its learning-free variant CMI-VLD$_{lf}$ on four LVLMs using greedy decoding. We present the results of the existing SOTA method, SID [9], as a reference.

Metric	Method	LLaVA-1.5	InstructBLIP	Shikra	LLaVA-NEXT
$C_S \downarrow$	SID	42.8	56.2	51.2	38.0
	CMI-VLD$_{lf}$	30.0	**40.4**	36.2	**26.6**
	CMI-VLD	**29.9**	43.2	**30.6**	27.2
$C_I \downarrow$	SID	12.1	15.8	13.6	8.9
	CMI-VLD$_{lf}$	9.0	**11.8**	10.2	**6.5**
	CMI-VLD	**8.9**	12.9	**8.9**	6.8
TPS $\uparrow$	SID	8.76	11.70	3.85	15.71
	CMI-VLD$_{lf}$	2.45	2.41	1.05	2.35
	CMI-VLD	**8.96**	**11.86**	**4.29**	**16.52**

Investigation of the learning-free variant. Initially, we proposed learning a purifier to reduce the computational overhead incurred by manual token selection. To validate this design, we implement a learning-free variant CMI-VLD$_{lf}$, which selects tokens by manually computing our derived score in Eq. (7) at each step, with all other settings unchanged.

Table 7 shows that both variants of our method significantly mitigate hallucination compared to the SOTA baseline, validating the effectiveness of our objective function derived from C-PMI. However, manual token selection incurs substantial latency due to repeated score computations at each decoding step, limiting its practicality in real-world applications. In contrast, our learned purifier efficiently selects informative tokens with nearly 4× faster inference than CMI-VLD$_{lf}$ while preserving strong effectiveness, exhibiting an excellent trade-off between performance and efficiency.

D Model Architecture of Visual Token Purifier

We provide the detailed purifier architecture as follows. Notably, this learnable network contains fewer than 1% of the LVLM's total parameters, hence introducing only marginal computation overheads.

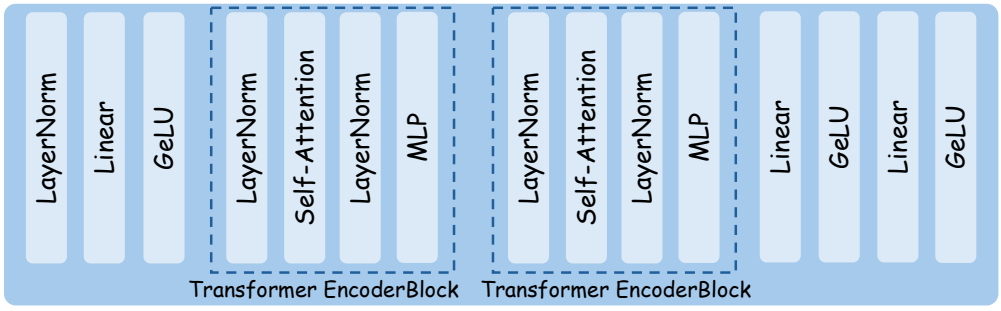


Figure 8: Overview of the architecture of the visual token purifier.

E Limitations

Despite the promising performance of LVLMs, the proposed method still faces several limitations. First, the introduced visual purifier introduces additional computational overhead for purifier training. Second, when the LVLM generates very long responses, the efficiency gain from the removal of visual tokens may become less significant. Besides, the performance of the LVLM is highly sensitive to the retaining ratio, and the optimal ratio may vary at each decoding step. However, our method adopts a fixed mask rate throughout the generation process. Future work could better address this limitation by exploring more advanced purification strategies with adaptive retaining ratios.

F Visualization Results

To visually demonstrate the effectiveness of our approach, we present generated captions using the prompt "`Please describe this image in detail.`" from our method and baseline approaches on the MSCOCO dataset in Fig. 9 and 10. Qualitative results reveal the effectiveness of our method.

Instruction

(max new tokens=512)

Please describe this image in detail.

Greedy: The image features a close-up view of a delicious sandwich with a slice of cucumber on top. The sandwich is cut in half, revealing the cucumber and other ingredients inside. The sandwich is placed on a dining table, accompanied by a bowl and a cup. The bowl is located on the left side of the table, while the cup is situated on the right side. The sandwich appears to be a delightful and healthy meal option.

SID: The image features a dining table with two small sandwiches on a piece of cardboard. Each sandwich is filled with tuna salad and topped with green sprouts or spinach leaves. A jar of greens, possibly parsley, is placed near the sandwiches. A knife can also be seen on the table, possibly used for spreading the tuna salad. The table setting is simple, with only the two sandwiches and the jar of greens occupying the space.

VASparse: The image features a close-up view of a delicious-looking sandwich, cut in half and placed on a dining table. The sandwich is filled with meat and accompanied by a leafy green garnish, adding a touch of freshness to the dish. The dining table is set with various items, including a wine glass, a cup, a fork, and a spoon. The wine glass is positioned on the left side of the table, while the cup is located near the center. The fork and spoon are placed on the right side of the table.

CMI-VLD: The image features two small sandwiches, each with a leaf sprouting from the top. The sandwiches are placed on a wooden table, and there is a jar nearby, possibly containing a spread or sauce. The sandwiches are cut in half, revealing the filling inside. The overall scene appears to be a casual and appetizing meal.

Instruction

(max new tokens=512)

Please describe this image in detail.

Greedy: The image features a delicious breakfast plate with a blueberry, strawberry, and banana pancake topped with whipped cream. The plate is placed on a dining table, and a cup of coffee is positioned next to it. The table is surrounded by chairs, with one chair on the left side and another on the right side. The arrangement creates a cozy and inviting atmosphere for enjoying a morning meal.

SID: The image showcases a plate with a breakfast pastry topped with strawberries and blueberries, sitting on a dining table. Beside the plate, there is a cup of coffee, enhancing the overall breakfast setting. There are two chairs visible in the scene; one is located to the right side of the table, while the other is situated towards the top-right corner. The breakfast setting appears to be well-prepared and inviting.

VASparse: The image features a delicious breakfast plate with a blueberry, strawberry, and banana pancake topped with whipped cream. The plate is placed on a dining table, accompanied by a cup of coffee. There are two chairs visible in the scene, one on the left side and another on the right side of the table. A fork can be seen on the right side of the plate, ready to be used to enjoy the scrumptious meal.

CMI-VLD: The image features a white plate with a delicious strawberry and blueberry topped pancake. The plate is placed on a dining table, and a coffee cup is situated next to it. The pancake is topped with fresh strawberries and blueberries, making it a visually appealing and appetizing breakfast or snack.

Figure 9: Visualization results comparing our CMI-VLD and other methods with LLaVA-1.5 backbone. Hallucinations are marked in red.

Instruction

(max new tokens=512)

Please describe this image in detail.

Greedy: The image features a train traveling over a bridge that spans a river. The train is positioned in the middle of the scene, with its length visible as it crosses the bridge. The bridge itself is made of concrete and appears to be a part of a larger infrastructure. In addition to the train, there are several birds flying in the sky, adding a sense of liveliness to the scene. The birds are scattered throughout the image, with some closer to the train and others further away. The combination of the train, the bridge, and the birds creates a dynamic and lively atmosphere.

SID: The scene features a train crossing over a red brick bridge that spans a body of water. The train is composed of two long silver and blue cars, moving on the bridge. The bridge itself has a large pipe on top, adding a distinctive architectural element to the scene. In addition to the train, there are three birds visible in the sky, possibly flying or soaring above the water. The overall setting gives a sense of tranquility, with the train gently traveling over the waterway.

VASparse: The image depicts a train traveling over a bridge that spans a river. The train is positioned in the middle of the scene, with its length visible as it crosses the bridge. The bridge itself is made of concrete and appears to be a part of a larger infrastructure. In addition to the train, there are a few people scattered throughout the scene, possibly enjoying the view or waiting for the train to pass. Some of them are closer to the train, while others are further away, near the edges of the image.

CMI-VLD: The image features a train traveling over a red bridge that spans a body of water. The train is blue and white, and it appears to be a commuter train. The bridge is made of concrete and provides a clear path for the train to cross the water. In the background, there are trees and a gas tower, adding to the scenery of the image. The combination of the train, the bridge, and the water creates a picturesque scene.

Instruction

(max new tokens=512)

Please describe this image in detail.

Greedy: The image features a kitchen with a sink and a refrigerator on the left side. Above the sink, there is a window with a projection screen, which is displaying a picture of a house. The kitchen is equipped with a microwave and an oven, both located on the countertop. There are several bottles placed around the kitchen, with one near the sink, another on the countertop, and a third one further away. A bowl can also be seen on the countertop, and a vase is placed near the sink. The kitchen appears to be well-equipped and organized, with a focus on the projection screen and the displayed image of the house.

SID: The image shows a kitchen with a stainless steel sink, a black refrigerator, and black cabinets. There are various items in the kitchen, including several bottles and cups placed around the countertop, a vase, a bowl, and a spoon. A TV screen is mounted in the kitchen, likely showing the view of the house outside. The room is brightly lit, giving a warm atmosphere to the space. In addition to the main appliances, there is a potted plant placed in the kitchen, adding a touch of greenery to the room.

VASparse: The image depicts a kitchen with a stainless steel sink, a refrigerator on the left side, and a microwave on the countertop. Above the sink, there is a projection screen displaying an image of a house. The kitchen is equipped with various appliances, including a toaster, a bottle, a bowl, and a spoon. There are two people in the scene, one standing near the sink and the other closer to the refrigerator. The kitchen appears to be well-organized and clean, with a focus on functionality and aesthetics.

CMI-VLD: The image features a kitchen with black cabinets and a sink. Above the sink, there is a window with a projection screen, which is displaying a house image. The kitchen also has a refrigerator on the left side and a bottle placed on the countertop. The overall atmosphere of the kitchen is cozy and inviting.

Figure 10: Visualization results comparing our CMI-VLD and other methods with LLaVA-1.5 backbone. Hallucinations are marked in red.