Mixture of Decoding: An Attention-Inspired Adaptive Decoding Strategy to Mitigate Hallucinations in Large Vision-Language Models

Xinlong Chen[1,2*], **Yuanxing Zhang**[3], **Qiang Liu**[1,2†], **Junfei Wu**[1,2], **Fuzheng Zhang**[3], **Tieniu Tan**[1,2,4]
[1]New Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Kuaishou Technology [4]Nanjing University

Abstract

Large Vision-Language Models (LVLMs) have exhibited impressive capabilities across various visual tasks, yet they remain hindered by the persistent challenge of hallucinations. To address this critical issue, we propose Mixture of Decoding (MoD), a novel approach for hallucination mitigation that dynamically adapts decoding strategies by evaluating the correctness of the model's attention on image tokens. Specifically, MoD measures the consistency between outputs generated from the original image tokens and those derived from the model's attended image tokens, to distinguish the correctness aforementioned. If the outputs are consistent, indicating correct attention, MoD employs a complementary strategy to amplify critical information. Conversely, if the outputs are inconsistent, suggesting erroneous attention, MoD utilizes a contrastive strategy to suppress misleading information. Extensive experiments demonstrate that MoD significantly outperforms existing decoding methods across multiple mainstream benchmarks, effectively mitigating hallucinations in LVLMs. The code is available at https://github.com/xlchen0205/MoD.

1 Introduction

Recently, Large Language Models (LLMs) (Touvron et al., 2023; Bai et al., 2023a) have achieved remarkable success, which has spurred significant research interest in extending them into Large Vision-Language Models (LVLMs) (Liu et al., 2024b; Bai et al., 2023b). These LVLMs demonstrate exceptional performance across various visual tasks. However, despite their impressive versatility, LVLMs are confronted with a critical limitation known as "hallucination" (Zhou et al., 2023; Bai et al., 2024; Rawte et al., 2023; Huang et al., 2023), a phenomenon in which models generate text that appears semantically coherent, but contains content that either contradicts or lacks grounding in the provided visual information. This fundamental flaw significantly compromises the reliability of LVLMs as trustworthy AI assistants in real-world applications (Wang et al., 2023b; He et al., 2023), particularly in high-stakes domains such as autonomous driving systems (Chen et al., 2024a; Tian et al., 2024), where inaccurate interpretation of visual data could lead to severe consequences.

A prominent approach to mitigating hallucinations in LVLMs is developing training-free decoding strategies, particularly contrastive decoding. This method works by contrasting the original logits for next-token prediction with those generated under hallucination scenarios, thereby producing more reliable outputs. Specifically, VCD (Leng et al., 2024) mitigates the language prior bias inherent in the LLM backbone by contrasting the original logits with those generated from Gaussian-noised image. Similarly, M3ID (Favero et al., 2024) reduces language prior bias by contrasting the original logits with those derived from pure text inputs, strengthening the interaction between visual inputs and model outputs. AvisC (Woo et al., 2024) proposes a novel perspective, suggesting that certain image tokens with excessively high attention weights may trigger hallucinations. Thus, it selects high-attention image tokens from specific layers to obtain hallucination logits for contrastive decoding.

Despite these advancements, existing methods still exhibit notable limitations. VCD and M3ID primarily attribute hallucinations to language prior bias, neglecting the potential influence of visual inputs, such as spurious correlations (Chen et al., 2024b). Although AvisC considers the model's attention distribution over the input image, it may erroneously weaken the role of high-attention image tokens when the model has already accurately focused on relevant information, leading to unreliable contrastive results. These limitations underscore

*Work done during an internship at Kuaishou Technology.
†Corresponding author: qiang.liu@nlpr.ia.ac.cn

Findings of the Association for Computational Linguistics: ACL 2025, pages 8525–8542
July 27 - August 1, 2025

a critical gap in current methods: the insufficient consideration of both (1) the influence of visual inputs on hallucination and (2) the variability of the attention distribution, which may be either right or wrong, ultimately restricting their performance in complex scenarios.

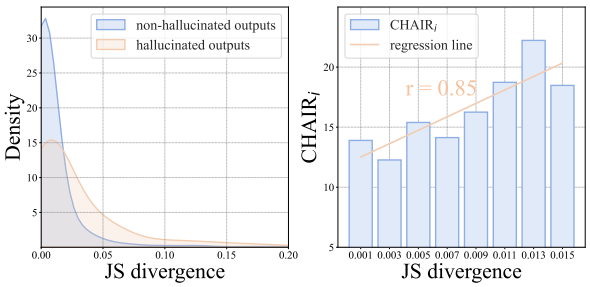


Figure 1: **The JS divergence excels at distinguishing hallucinated from non-hallucinated outputs**, as evidenced in both discriminative tasks (POPE, left) and generative tasks (CHAIR, right). In the left figure, non-hallucinated outputs are primarily concentrated in regions with lower JS divergence values, while hallucinated outputs exhibit a pronounced long-tail distribution. In the right figure, there is a significant positive correlation between the CHAIR$_i$ metric and JS divergence, with a Pearson correlation coefficient of 0.85 ($p < 0.01$).

The key to addressing these limitations lies in assessing the correctness of the model's attention on image tokens and dynamically adjusting the decoding strategy accordingly. If the model's attention is correct, indicating that the attended image tokens are relevant and beneficial, amplifying their information can enhance the output probability of proper tokens; conversely, if the model's attention is incorrect, suggesting that the attended image tokens are irrelevant or misleading, suppressing erroneous information is crucial to counteract hallucinations. Inspired by methods that judge correctness based on consistency (Manakul et al., 2023; Liu et al., 2025), we observe that measuring the consistency between the original output logits and the logits derived from the model's attended image tokens using Jensen-Shannon (JS) divergence, can effectively distinguish hallucinated outputs from non-hallucinated ones (as illustrated in Fig. 1). This approach enables us to determine the correctness of the model's attention on image tokens and adaptively select decoding strategies to effectively mitigate hallucinations.

Building on these insights, we propose **Mixture of Decoding (MoD)**, a method that adaptively selects decoding strategies by evaluating the consistency between the original output and the output derived from the model's attended image tokens. Specifically, if the outputs are consistent, it indicates that the model's attention on image tokens is correct, and we complement the two to amplify critical information. Conversely, if the outputs are inconsistent, it suggests that the model's attention on image tokens is incorrect, and we contrast the two to suppress misleading information derived from erroneous attention. Extensive experiments on multiple mainstream benchmarks demonstrate that MoD significantly outperforms existing methods in mitigating hallucinations.

The main contributions of this paper are summarized as follows:

- We demonstrate that the consistency between the outputs derived from the original image tokens and the model's attended image tokens provides a robust criterion for distinguishing hallucinations.
- We propose MoD, a novel method that dynamically selects decoding strategies by evaluating the aforementioned consistency, significantly reducing hallucinations in LVLMs.
- Our method achieves state-of-the-art performance across multiple prominent benchmarks, highlighting the efficacy and superiority of MoD.

2 Related Work

Large vision-language models. Inspired by the success of LLMs (Touvron et al., 2023; Bai et al., 2023a), recent research has extended their capabilities into the multimodal domain, giving rise to the developments of LVLMs (Liu et al., 2024b; Bai et al., 2023b). By integrating textual instructions with visual inputs, LVLMs are capable of understanding and generating diverse content in a more comprehensive manner. Typically, LVLMs undergo two training stages: feature alignment pretraining and instruction-based fine-tuning. Additionally, recent efforts have attempted to enhance the alignment of model responses with human values through reinforcement learning from human feedback (RLHF) (Sun et al., 2023) and preference fine-tuning (Zhou et al., 2024). Despite these achievements, existing LVLMs still struggle with significant hallucination issues, which severely undermine their reliability and stability in practical applications. The MoD approach proposed in this paper offers a promising solution to effectively mitigate hallucinations in LVLMs.

Hallucination mitigation methods in LVLMs. Hallucination mitigation methods can be systemat-

ically categorized into three major paradigms: pre-inference intervention, in-inference regulation, and post-inference correction. In terms of pre-inference intervention, researchers primarily employ specifically curated anti-hallucination data to perform supervised fine-tuning (SFT) on LVLMs (Jiang et al., 2024; Yu et al., 2024). In-inference regulation strategies involve adjustments to the internal components of LVLMs, including enhancing or suppressing attention heads associated with facts or errors (Li et al., 2024; Yuan et al., 2024), as well as various decoding strategies (Leng et al., 2024; Favero et al., 2024; Woo et al., 2024). Post-inference correction methods mainly detect hallucinations by designing verification questions and retrieve external knowledge for correction (Varshney et al., 2023; Wu et al., 2024; Yin et al., 2024; Luo et al., 2023; Dhuliawala et al., 2023). However, existing methods exhibit significant limitations. Pre-inference intervention approaches rely heavily on large-scale, manually constructed or AI-generated training data, resulting in high costs. In-inference regulation methods, which involve adjusting internal components of the model, require extensive testing on large datasets to identify specific components that need regulation, and may inadvertently impair other capabilities of LVLMs. Post-inference correction methods necessitate multiple inferences to detect hallucinations and depend on external knowledge or tools for correction, significantly increasing inference time. In contrast, contrastive decoding offers a more efficient solution for hallucination mitigation, as it requires no additional training, no large-scale data testing, and no extensive repeated sampling.

Contrastive decoding. As a hallucination mitigation method under the in-inference regulation paradigm, contrastive decoding (Li et al., 2022) was initially proposed to leverage the contrast between an expert model and an amateur model, generating outputs superior to those of the expert model alone. Recently, this approach has been extended, with researchers employing various designs to replace the amateur model, intentionally inducing hallucinations and contrasting them with the original outputs to obtain hallucination-free results. For instance, VCD (Leng et al., 2024) amplifies language prior bias by adding Gaussian noise to the input image, while M3ID (Favero et al., 2024) removes the image input entirely and further amplifies language prior bias using pure text inputs. AvisC (Woo et al., 2024) posits that image patches with excessively high attention weights can trigger hallucinations, while SID (Huo et al., 2024) takes the opposite stance; ICD (Wang et al., 2024) introduces a creative method by designing diverse negative prompts to induce hallucinations; DoLa (Chuang et al., 2023) suggests that early-layer information, being less mature, can be used to induce hallucinations; ID (Kim et al., 2024) employs noisy instructions to induce hallucinations; and IBD (Zhu et al., 2024) fine-tunes an image-biased LVLM by amplifying attention to visual tokens and contrasts it with the original LVLM. DeGF (Zhang et al., 2025) proposes a novel approach that first generates an image from text based on the original output, then adaptively selects decoding strategies according to the consistency between the output derived from the generated image and the original output, achieving remarkable hallucination detection performance. Although these methods exhibit certain effectiveness, they often fail to adequately address the impact of visual inputs on hallucinations or overlook the variability of attention distribution, which can be either correct or incorrect. To overcome these limitations, we propose MoD, which adaptively selects decoding strategies based on the consistency between outputs derived from the model's attended image tokens and its original outputs, offering a more promising solution for hallucination mitigation.

3 Method

3.1 Preliminary: Attention Formulation

In the transformer architecture, the attention mechanism plays a pivotal role in the model's decision-making process. Let's begin by defining an embedding sequence $X \in \mathbb{R}^{N \times d}$, where N is the total number of tokens, comprising N_I image tokens and N_T text tokens such that $N = N_I + N_T$, and d represents the dimension of hidden states. This sequence is projected into three distinct matrices: the query Q, the key K, and the value V, through three learnable linear transformations: W_Q, W_K, and W_V. The attention weight matrix A is then computed as follows:

$$A = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right). \tag{1}$$

The output of the attention mechanism O_A is derived by $O_A = A \cdot V$. Let $A_l \in \mathbb{R}^{B \times H \times N \times N}$ represent the attention weight matrix of layer l in the LLM backbone consisting of L layers of an

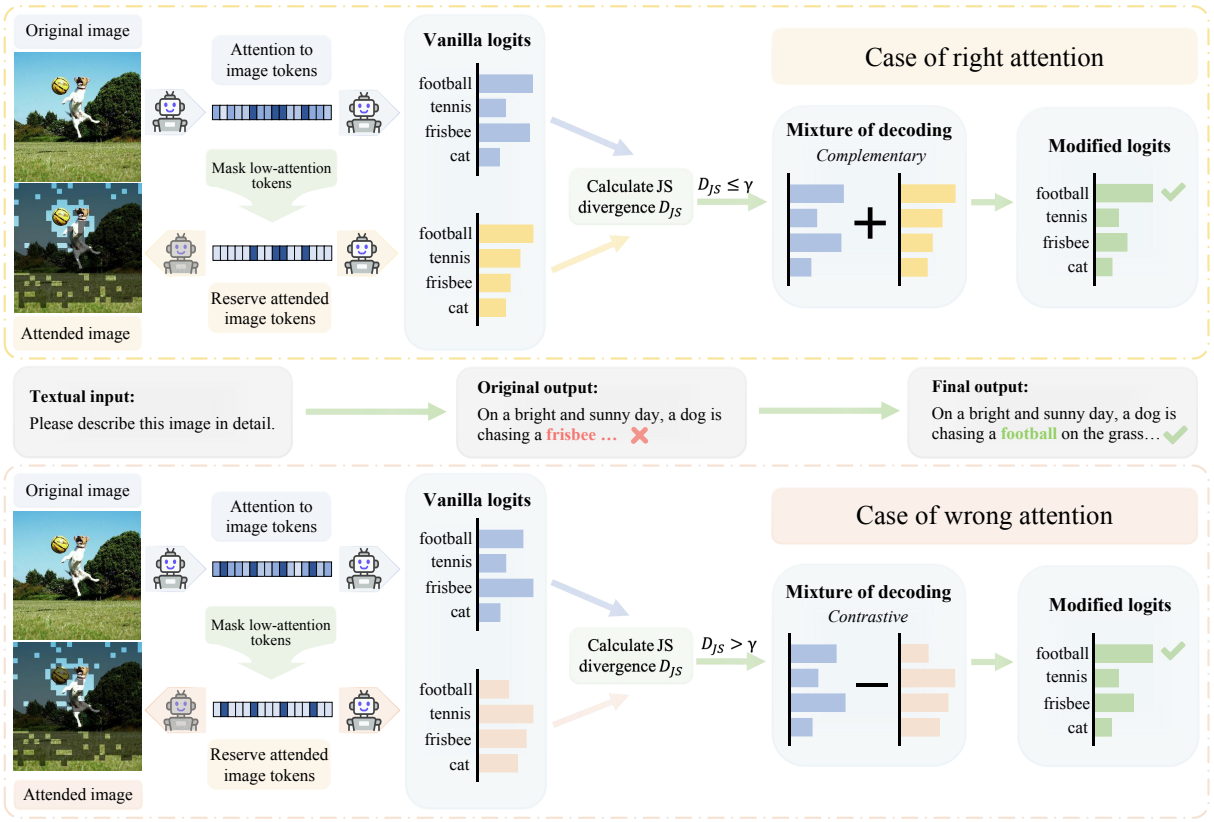


Figure 2: **Overview of our proposed MoD.** MoD involves three key steps: (1) extracting the model's attended image tokens while masking the others; (2) generating vanilla output logits from both original image tokens and masked image tokens; and (3) computing the JS divergence between the two logit distributions to assess the correctness of the model's attention. Based on this evaluation, MoD adaptively adopts either complementary or contrastive decoding strategies to produce hallucination-free outputs. The upper and lower panels illustrate cases of right and wrong attention, respectively. The corresponding attended image, visualized using LLaVA-1.5, is displayed in the lower-left corner of each panel.

LVLM, where B and H denote the batch size and the number of attention heads, respectively.

3.2 Extraction of Attended Image Tokens

Due to the auto-regressive nature of LLMs, it can be posited that the final token of the input sequence encapsulates the model's comprehensive understanding of the input query, signifying its readiness to formulate a response. Therefore, we utilize its attention to the image tokens, denoted as $A^I \in \mathbb{R}^{B \times N_I}$, to identify the specific image tokens that the model has attended to. Specifically, we compute the average attention weight across all layers and attention heads to integrate the model's interpretation of the input across various levels and dimensions of abstraction, formalized as:

$$A^I = \frac{1}{L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} A_l\left[\cdot\,, h\,, -1\,, IDX^I\right], \quad (2)$$

where IDX^I represents the indices of image tokens within the input sequence. To address the variability in the number of image tokens and the discrepancies in the overall attention weights assigned to the image across different models, we select the top λ proportion of image tokens with the highest attention weights as the model's most attended ones, with their indices denoted as IDX^I_{att}:

$$IDX^I_{att} = \text{argtopk}_{i \in IDX^I}(A^I\left[\cdot\,, i\right], \lambda), \quad (3)$$

Given the original image tokens v and indices of the model's attended image tokens IDX^I_{att}, we zero out the remaining image tokens, whose indices are $IDX^I \setminus IDX^I_{att}$, resulting in the model's attended image tokens v_{att}.

3.3 Mixture of Decoding

The decoding process is conducted within the LLM backbone of the LVLM, parameterized by θ. Given the image tokens v, text input tokens x, and previous output tokens $y_{<t}$, the probability distribution

for the t-th token y_t is expressed as:

$$y_t \sim p_\theta(y_t \mid v, x, y_{<t}) \propto \exp \text{logit}_\theta(y_t \mid v, x, y_{<t}). \tag{4}$$

Similarly, the probability distribution for y_t can also be derived based on the model's attended image tokens v_{att}, text input tokens x, and previous output tokens $y_{<t}$, denoted as $p_\theta(y_t \mid v_{att}, x, y_{<t})$. Subsequently, we define $d(v, v_{att})$ as the JS divergence between the output probability distributions of y_t obtained from the two different sets of image tokens v and v_{att}:

$$d(v, v_{att}) = D_{JS}\big[p_\theta(y_t \mid v, x, y_{<t}) \parallel p_\theta(y_t \mid v_{att}, x, y_{<t})\big]. \tag{5}$$

If the output based on v_{att} aligns with the original output derived from v, it indicates that the model's attention over image tokens is right, as illustrated in the upper part of Fig. 2. In this scenario, the two outputs are complemented to enhance the output probability of proper tokens. Conversely, if the outputs are inconsistent, it suggests that the model's attention over image tokens is wrong, as depicted in the lower part of Fig. 2. In this case, the two outputs are contrasted to counteract the hallucination caused by the erroneous attention. Thus, the MoD strategy is formulated as follows:

$$y_t \sim p_\theta(y_t \mid v, v_{att}, x, y_{<t}) = \begin{cases} \text{softmax}\big[\text{logit}_\theta(y_t \mid v, x, y_{<t}) + \\ \alpha_1 \cdot \text{logit}_\theta(y_t \mid v_{att}, x, y_{<t})\big], \text{if } d(v, v_{att}) \le \gamma; \\ \text{softmax}\big[(1+\alpha_2) \cdot \text{logit}_\theta(y_t \mid v, x, y_{<t}) - \\ \alpha_2 \cdot \text{logit}_\theta(y_t \mid v_{att}, x, y_{<t})\big], \text{if } d(v, v_{att}) > \gamma. \end{cases} \tag{6}$$

Here, α_1 and α_2 denote the hyperparameters for the complement and contrast operations, respectively, and γ represents the JS divergence threshold, which is used to determine the consistency between the two outputs.

4 Experiments

4.1 Experimental Settings

Benchmarks and metrics. (1) **POPE** (Li et al., 2023) is a widely used benchmark designed to evaluate object hallucination. It employs yes/no questions formulated as "Is there an {object} in the image?" to query the model about the presence of specific objects in an image. The evaluation metrics include Accuracy, Precision, Recall, and F1 score. (2) **MME** (Fu et al., 2023) is a more challenging benchmark for hallucination assessment. It designs a pair of similar questions with answers "yes" and "no", respectively, for each image. In addition to question-level accuracy (Acc), Acc+ is introduced to represent image-level accuracy. An image is deemed to be fully understood only if the model answers both associated questions correctly. The overall performance is quantified by the MME Score, which is calculated as the sum of Acc and Acc+. (3) **CHAIR** (Rohrbach et al., 2018) evaluates object hallucination in image captioning tasks. In this evaluation framework, LVLMs are asked to generate descriptive captions for a randomly selected subset of 500 images from the MS-COCO validation set. Specifically, CHAIR quantifies hallucination by calculating the proportion of objects mentioned in the model's caption that do not exist in the ground truth. The main metrics comprise CHAIR_i and CHAIR_s, which assess hallucination at the instance level and the sentence level, respectively. (4) **AMBER** (Wang et al., 2023a) is a more comprehensive benchmark that evaluates object hallucination, attribute hallucination, and relation hallucination in both discriminative and generative tasks. For discriminative tasks, evaluation metrics include Accuracy, Precision, Recall, and F1 score. For generative tasks, the metrics include CHAIR_i, Cover, Hal, and Cog (see Appendix B.1 for details). Additionally, the AMBER Score is proposed to provide a unified evaluation, calculated as:

$$\text{AMBER Score} = \frac{1}{2} \times (1 - \text{CHAIR}_i + \text{F1}). \tag{7}$$

Models. The performance of MoD is evaluated on the 7B versions of LLaVA-1.5 (Liu et al., 2024b), Qwen-VL (Bai et al., 2023b), and LLaVA-NEXT (Liu et al., 2024a), respectively. LLaVA-1.5 employs a two-layer MLP to align vision and language modalities, while Qwen-VL uses a single-layer cross-attention module to compress and align visual features to the language modality. LLaVA-NEXT further enhances LLaVA-1.5 by introducing the "anyres" mechanism, which increases the input image resolution by $4\times$ to enable fine-grained visual understanding. Notably, MoD is model-agnostic and can be integrated into diverse LVLM architectures.

Baselines. We first adopt sampling decoding as the most basic baseline, where the next token is sampled directly from the output probability distribution. Additionally, we compare our method with three contrastive decoding strategies: VCD (Leng et al., 2024), M3ID (Favero et al., 2024), and

AvisC (Woo et al., 2024). To ensure a fair comparison, we report the results of these methods based on our reproduction using their official codebases within our own experimental environment.

Implementation Details. In contrast to many existing decoding methods that necessitate extensive task- and model-specific hyperparameter tuning, MoD demonstrates robust performance across diverse tasks and models using a shared set of hyperparameters, with low sensitivity to their selection (see Sec. 4.3 for detailed analysis). Specifically, for the extraction of the model's attended image tokens (Eq. 3), we set $\lambda = 0.2$; for the decoding process (Eq. 6), we set $\alpha_1 = 4$, $\alpha_2 = 1$, and $\gamma = 0.05$. All experiments are conducted on NVIDIA RTX 3090 GPUs with 24GB of memory, utilizing a fixed random seed of 42 for reproducibility.

4.2 Experimental Results

Results on POPE. Tab. 1 presents the performance of various decoding methods on the POPE benchmark based on the MS-COCO dataset. The complete experimental results are provided in Appendix B.3. As shown, MoD achieves the best performance across all three evaluation settings, outperforming the second-best method by up to 3.1 points in Accuracy and 4.1 points in F1 score. These results demonstrate that MoD enhances the model's ability to perceive fine-grained object details in images, effectively mitigating object hallucinations. Additionally, MoD exhibits remarkable superiority in Precision, surpassing other methods by up to 6.8 points, indicating fewer false positive samples. This suggests that the model becomes more conservative when generating positive answers, effectively suppressing the tendency of existing LVLMs to answer "Yes" without adequately considering the question itself.

Results on MME. Following prior studies, we evaluate object and attribute hallucination on four subsets of the MME benchmark, as detailed in Tab. 2. Overall, MoD shows significant improvements over the second-best method, with notable gains in MME Score of 41.6 points for LLaVA-1.5, 20.0 points for Qwen-VL, and 40.0 points for LLaVA-NEXT. However, there is a slight decline in the "position" subset for LLaVA-1.5 and Qwen-VL. We hypothesize that this is due to the masking of low-attention image tokens, which employs a zeroing-out approach, potentially reducing the model's sensitivity to relative positional information. This issue might be addressed by refining the masking strategy, such as incorporating pooling mechanisms to allow the model to retain partial awareness of low-attention image tokens. We leave this exploration for future work. Notably, this decline is not observed in LLaVA-NEXT, which we attribute to its "anyres" mechanism, where the masked local and global image tokens complement each other, thereby preserving relative positional information. Despite this, MoD consistently outperforms other decoding methods on the remaining three subsets and achieves the highest total MME Score, demonstrating its overall effectiveness in mitigating hallucinations.

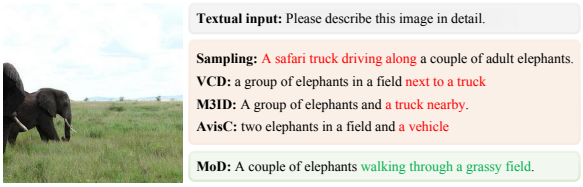


Figure 3: **Case study of generative tasks using Qwen-VL.** We compare responses generated by sampling, VCD, M3ID, AvisC, and our proposed MoD. Hallucinated content is highlighted in red, while more detailed and accurate content is marked in green.

Results on CHAIR. In addition to its strong performance in discriminative tasks, MoD also exhibits exceptional capabilities in generative tasks. As evidenced by the quantitative results in Tab. 3, MoD achieves the best performance across both sentence-level (CHAIR_s) and instance-level (CHAIR_i) hallucination evaluation metrics. Moreover, MoD also excels in Recall, which measures the completeness of generated captions. Notably, despite using the prompt "Please describe this image in detail", Qwen-VL tends to produce concise outputs, resulting in significantly lower absolute values for all metrics. However, MoD still manages to reduce hallucination in Qwen-VL while simultaneously improving its output completeness. Qualitative results presented in Fig. 3 further illustrate this point, demonstrating the effectiveness of MoD regardless of the caption length. Additionally, while M3ID intensifies hallucinations in LLaVA-1.5, and both VCD and AvisC exacerbate hallucinations in LLaVA-NEXT, MoD consistently mitigates hallucinations across all models, highlighting its robustness and model-agnostic characteristics.

Results on AMBER. The AMBER benchmark evaluates both discriminative and generative tasks, with quantitative results shown in Tab. 4. In discriminative tasks, MoD maintains the leading per-

Setting	Method	LLaVA-v1.5				Qwen-VL				LLaVA-NEXT			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
random	sampling	83.8	82.4	86.1	84.2	84.9	96.0	72.9	82.9	84.4	94.7	72.8	82.3
	VCD	85.0	82.7	86.1	84.2	85.5	96.0	71.1	83.6	86.0	96.5	74.8	84.3
	M3ID	86.1	83.2	86.8	85.0	85.3	95.1	74.2	83.4	85.5	96.3	73.9	83.6
	AvisC	82.3	78.4	89.3	83.5	82.9	96.2	68.5	80.0	85.2	98.6	71.3	82.8
	MoD (Ours)	**89.2**	90.0	88.2	**89.1**	**86.0**	97.0	74.3	**84.1**	**86.6**	97.7	74.9	**84.8**
popular	sampling	82.0	79.7	85.9	82.6	84.0	94.7	72.1	81.9	83.2	90.9	73.8	81.5
	VCD	82.1	78.5	88.3	83.2	84.9	94.5	74.9	83.6	84.5	92.9	74.8	82.9
	M3ID	82.8	80.1	88.5	84.1	84.2	94.1	73.8	82.7	84.2	93.1	73.9	82.4
	AvisC	78.2	72.7	90.3	80.5	82.8	95.5	68.9	80.1	83.9	94.7	71.7	81.6
	MoD (Ours)	**85.7**	84.1	88.1	**86.1**	**85.6**	96.3	74.0	**83.7**	**85.5**	95.1	74.9	**83.8**
adversarial	sampling	75.8	71.3	86.3	78.1	82.1	90.0	72.3	80.2	79.5	84.1	72.9	78.1
	VCD	76.3	71.5	87.3	78.7	84.0	90.6	74.9	82.0	80.9	85.2	74.8	79.7
	M3ID	77.1	71.8	87.6	78.9	83.2	90.4	73.1	80.8	80.6	85.4	73.9	79.2
	AvisC	74.2	68.4	89.9	77.7	81.2	91.9	68.5	78.5	81.8	91.3	71.5	80.2
	MoD (Ours)	**79.7**	75.4	88.2	**81.3**	**84.0**	92.4	74.2	**82.3**	**82.4**	88.3	74.8	**81.0**

Table 1: **Results on POPE benchmark.** "Acc", "Pre", "Rec", and "F1" stand for Accuracy, Precision, Recall, and F1 score, respectively. The reported results are derived from the MS-COCO dataset.

Model	Method	Object-level		Attribute-level		Total
		Existence	Count	Position	Color	
LLaVA-v1.5	sampling	170.0	103.3	108.3	128.3	510.0
	VCD	180.0	110.0	108.3	133.3	531.7
	M3ID	185.0	118.3	121.7	128.3	553.3
	AvisC	195.0	116.7	131.7	153.3	596.7
	MoD (Ours)	195.0	141.7	126.7	175.0	**638.3**
Qwen-VL	sampling	160.0	143.3	113.3	165.0	581.7
	VCD	165.0	140.0	113.3	175.0	593.3
	M3ID	165.0	143.3	103.3	175.0	586.7
	AvisC	160.0	145.0	113.3	160.0	578.3
	MoD (Ours)	170.0	160.0	103.3	180.0	**613.3**
LLaVA-NEXT	sampling	175.0	143.3	131.7	145.0	595.0
	VCD	190.0	145.0	116.7	160.0	611.7
	M3ID	195.0	145.0	103.3	165.0	608.3
	AvisC	195.0	160.0	108.3	150.0	613.3
	MoD (Ours)	195.0	160.0	133.3	165.0	**653.3**

Table 2: **Results on MME benchmark.** The performance is measured by MME Score. The "Total" column represents the sum of four individual results in each row.

Model	Method	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall ↑	Length
LLaVA-v1.5	sampling	52.8	15.9	77.3	93.4
	VCD	51.0	14.9	77.2	101.9
	M3ID	56.2	17.0	**79.3**	97.1
	AvisC	44.0	13.7	72.9	89.8
	MoD (Ours)	**42.6**	**12.4**	78.9	97.6
Qwen-VL	sampling	2.8	3.0	31.0	5.3
	VCD	1.4	1.2	30.8	4.0
	M3ID	1.7	1.3	31.8	3.4
	AvisC	1.6	1.6	32.0	4.4
	MoD (Ours)	**0.8**	**1.0**	**32.1**	3.8
LLaVA-NEXT	sampling	35.8	12.0	59.5	179.0
	VCD	40.2	10.7	**62.1**	171.2
	M3ID	35.2	10.3	61.4	152.7
	AvisC	40.4	12.4	60.0	183.8
	MoD (Ours)	**33.6**	**9.6**	61.4	174.9

Table 3: **Results on CHAIR benchmark.** Lower $CHAIR_s$ and $CHAIR_i$, along with higher Recall, correspond to better performance.

formance in Accuracy and F1 score, achieving average improvements of 4.2 and 3.7 points, respectively, across the three evaluated models. In generative tasks, MoD significantly outperforms other methods in the two hallucination evaluation metrics, $CHAIR_i$ and Hal, with notable average performance gains of 2.9 and 2.1 points, respectively, while maintaining excellent coverage. Overall, MoD attains the highest AMBER score across all evaluated models, with improvements of 2.2 points for LLaVA-1.5, 0.7 points for Qwen-VL, and 2.6 points for LLaVA-NEXT compared to the second-best method, showcasing its versatility in handling both discriminative and generative tasks. Additional qualitative results, provided in Appendix C, further illustrate that MoD not only effectively mitigates hallucinations but also enriches the detail and depth of generated image captions.

4.3 Ablation Studies

Comparison with individual methods. To evaluate the performance gains brought by MoD compared to employing either the complementary method or the contrastive method independently, which corresponds to setting the consistency threshold γ in Eq. 6 to 1 and 0, respectively, we conduct ablation studies on the MME benchmark. The results, as detailed in Tab 5, demonstrate that MoD effectively integrates the strengths of both individual methods, achieving maximum performance improvements of 23.3 points over the complementary method alone and 20.0 points over the contrastive

Model	Method	Discriminative				Generative				AMBER
		Acc	Pre	Rec	F1	CHAIR$_i$ ↓	Cover ↑	Hal ↓	Cog*	Score
LLaVA-v1.5	sampling	67.0	85.2	60.9	71.0	12.0	50.3	51.0	4.6	79.5
	VCD	67.3	86.1	60.5	71.1	10.0	**51.2**	43.6	4.3	80.6
	M3ID	67.3	86.5	60.1	70.9	9.1	49.4	42.8	4.4	80.9
	AvisC	70.7	85.5	65.6	74.2	11.8	50.4	51.2	5.0	81.2
	MoD (Ours)	**72.4**	91.5	64.0	**75.3**	**8.6**	50.7	**38.8**	4.7	**83.4**
Qwen-VL	sampling	82.9	88.0	85.9	86.9	4.8	31.3	9.2	0.3	91.1
	VCD	84.1	89.2	86.6	87.9	3.5	35.2	7.7	0.3	92.2
	M3ID	84.0	88.1	86.2	87.1	3.7	34.1	8.2	0.4	91.7
	AvisC	84.1	88.8	87.0	87.9	4.3	35.0	8.6	0.4	91.8
	MoD (Ours)	**84.2**	89.2	86.7	**87.9**	**2.2**	**36.5**	**4.5**	0.3	**92.9**
LLaVA-NEXT	sampling	72.9	82.4	75.2	78.6	12.0	56.5	59.6	5.1	83.3
	VCD	74.3	83.9	75.8	79.6	11.8	59.1	58.6	5.0	83.9
	M3ID	75.0	84.9	75.7	80.0	10.2	59.0	52.8	4.6	84.9
	AvisC	77.7	88.0	76.9	82.1	14.2	58.1	65.2	5.7	84.0
	MoD (Ours)	**78.9**	83.4	85.2	**84.3**	**9.4**	**61.6**	**50.8**	3.2	**87.5**

Table 4: **Results on AMBER benchmark.** In discriminative tasks, "Acc", "Pre", "Rec", and "F1" stand for Accuracy, Precision, Recall, and F1 score, respectively. Higher values for these metrics indicate superior performance. In generative tasks, lower CHAIR$_i$ and Hal, along with higher Cover, signify better performance. *Cog measures the extent to which hallucinations in LVLMs align with human cognition and is therefore not directly comparable.

Model	Method	Object-level		Attribute-level		Total
		Existence	Count	Position	Color	
LLaVA -v1.5	sampling	170.0	103.3	108.3	128.3	510.0
	complementary	195.0	143.3	116.7	160.0	615.0
	contrastive	190.0	145.0	120.0	165.0	620.0
	MoD (Ours)	195.0	141.7	126.7	175.0	**638.3**
Qwen -VL	sampling	160.0	143.3	113.3	165.0	581.7
	complementary	175.0	145.0	103.3	180.0	603.3
	contrastive	175.0	155.0	103.3	170.0	603.3
	MoD (Ours)	170.0	160.0	103.3	180.0	**613.3**
LLaVA -NEXT	sampling	175.0	143.3	131.7	145.0	595.0
	complementary	195.0	145.0	133.3	165.0	638.3
	contrastive	190.0	150.0	138.3	155.0	633.3
	MoD (Ours)	195.0	160.0	133.3	165.0	**653.3**

Table 5: **Ablation study of individual decoding methods.** The performance of sampling, complementary, and contrastive decoding methods are compared with MoD on MME benchmark across three LVLMs.

method alone. This underscores the effectiveness of MoD in implementing adaptive decoding strategies by assessing the correctness of the model's attention over image tokens.

Analysis of the consistency threshold γ. Fig. 4 presents a comprehensive evaluation of the performance of MoD across varying consistency thresholds on the MME benchmark. The experimental results indicate that MoD maintains a consistent performance advantage over a wide range of consistency thresholds from 0.02 to 0.08, significantly outperforming both complementary and contrastive methods when employed independently. Specifically, the average MME Score achieved by MoD exceeds that of the best individual method by 14.0 points for LLaVA-1.5, 8.6 points for Qwen-VL, and 10.0 points for LLaVA-NEXT, underscoring the robustness of MoD across varying consistency threshold configurations.

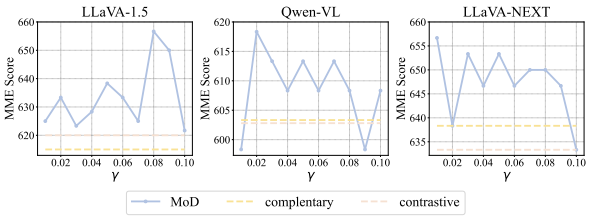


Figure 4: **Ablation study on the consistency threshold γ.** We vary γ from 0.01 to 0.10 with an increment of 0.01, and compare the results with those obtained using individual methods on MME benchmark. In MoD, γ is consistently set to 0.05 across all experiments.

We provide a detailed analysis of additional hyperparameters of MoD in Appendix B.4. Notably, MoD employs a shared set of hyperparameters across different tasks and model architectures, distinguishing it from most other decoding methods that necessitate meticulous hyperparameter tuning for specific scenarios. Furthermore, experimental results demonstrate that MoD exhibits low sensitivity to hyperparameter variations, further highlighting its practicality and adaptability.

5 Conclusion

In this paper, we propose Mixture of Decoding (MoD), an adaptive decoding strategy designed to mitigate hallucinations in LVLMs. By evaluating the consistency between the outputs generated from the original image tokens and those derived from the model's attended image tokens, MoD initially assesses the correctness of the model's attention over image tokens. Subsequently, informed by this evaluation, MoD dynamically selects either the complementary or contrastive decoding strategy to optimize the decoding process. A series of comprehensive experiments have been conducted to validate the effectiveness and robustness of MoD, highlighting its potential as a promising solution to enhance the accuracy and reliability of LVLMs across diverse applications.

Limitations

While MoD demonstrates promising results, we acknowledge several limitations of our approach.

First, similar to other decoding methods, MoD requires two forward passes, nearly doubling the computational latency compared to standard decoding. However, MoD achieves significantly better performance under the same inference cost.

Additionally, our current implementation employs a straightforward masking strategy that zeros out low-attention image tokens. Exploring more sophisticated masking strategies, such as pooling mechanisms, could potentially further enhance performance across a broader range of tasks. We leave this investigation for future work.

Ethical Considerations

Although MoD is designed to mitigate hallucinations and improve the reliability of LVLMs, it does not inherently address biases embedded in training data, which may lead to unfair or discriminatory outcomes. We emphasize the critical importance of prioritizing fairness and responsible deployment to ensure that LVLMs, augmented by MoD, contribute positively to society while minimizing potential ethical risks.

Acknowledgements

This work is jointly sponsored by National Natural Science Foundation of China (62236010, 62141608, 62206291).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. 2024a. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F Fouhey, and Joyce Chai. 2024b. Multi-object hallucination in vision-language models. *arXiv preprint arXiv:2407.06192*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023.

A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Taehyeon Kim, Joonkee Kim, Gihun Lee, and Se-Young Yun. 2024. Instructive decoding: Instruction-tuned large language models are self-refiner from noisy instructions. In *The Twelfth International Conference on Learning Representations*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Qiang Liu, Xinlong Chen, Yue Ding, Shizhen Xu, Shu Wu, and Liang Wang. 2025. Attention-guided self-reflection for zero-shot hallucination detection in large language models. *arXiv preprint arXiv:2501.09997*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

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

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023b. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

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

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2024. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953.

Hongbang Yuan, Pengfei Cao, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Whispers that shake foundations: Analyzing and mitigating false premise hallucinations in large language models. *arXiv preprint arXiv:2402.19103*.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. 2025. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

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

Appendix

A More Implementation Details

A.1 Adaptive Plausibility Constraints

Typically, current decoding methods employ adaptive plausibility constraints to calibrate the output distribution, truncating the generation of implausible outputs. The plausible vocabulary $V_{\text{head}}(y_{<t})$ is defined as:

$$V_{\text{head}}(y_{<t}) = \{y_t \in V : p_\theta(y_t \mid v, x, y_{<t}) \geq \beta \max_w p_\theta(w \mid v, x, y_{<t})\}, \tag{8}$$

where V is the whole vocabulary of the LVLM, and β controls the strength of truncation. Any potential output y_t not residing within $V_{\text{head}}(y_{<t})$ is assigned a probability of zero:

$$p_\theta(y_t \mid v, x, y_{<t}) = 0, \quad \text{if } y_t \notin V_{\text{head}}(y_{<t}). \tag{9}$$

A larger β indicates more aggressive truncation, preserving only tokens with higher output probabilities. In our MoD framework, we set $\beta = 0.5$ across all models and tasks.

A.2 General Decoding Parameters

To guarantee adequate generation diversity, we set $temperature = 1$ and $Top\text{-}p = 1$. Additionally, we configure $max\ new\ tokens = 1024$, enabling the model to autonomously determine the appropriate stopping point for text generation, which fully reflects the model's capabilities.

B More Experimental Details

B.1 Metrics

Metrics on MME. In the image set $\mathcal{I}$ of the MME benchmark, each image $i \in \mathcal{I}$ is paired with a set of two similar questions designed to elicit "yes" and "no" responses, denoted as $\{q_y^i, q_n^i\}$. To comprehensively evaluate the performance of models, two metrics are introduced: the question-level accuracy (Acc) and the image-level accuracy (Acc+). A correctly answered question contributes to Acc, while a correctly answered pair of questions for an image contributes to Acc+. Although Acc+ is a more rigorous metric, it provides a more comprehensive assessment of the model's ability to fully understand the images. The formulas for these metrics are defined as follows:

$$\text{Acc} = \Big\{ \sum_{i \in \mathcal{I}} \big[\text{LVLM}(i, q_y^i) = \text{"Yes"}\big] + \sum_{i \in \mathcal{I}} \big[\text{LVLM}(i, q_n^i) = \text{"No"}\big] \Big\} \,/\, \big(|\mathcal{I}| \times 2\big), \tag{10}$$

$$\text{Acc+} = \sum_{i \in \mathcal{I}} \big[\text{LVLM}(i, q_y^i) = \text{"Yes"} \wedge \text{LVLM}(i, q_n^i) = \text{"No"}\big] \,/\, |\mathcal{I}|, \tag{11}$$

where $\text{LVLM}(i, q_*^i)$ denotes the model's output when provided with image i and the corresponding question q_*^i. The reported MME Score is computed as the sum of Acc and Acc+.

Metrics on CHAIR. The CHAIR evaluation benchmark comprises two key metrics: CHAIR_i and CHAIR_s, which measure hallucinations in image captions at the instance level and sentence level, respectively. Specifically, the instance-level metric CHAIR_i denotes the proportion of hallucinated objects relative to all mentioned objects in the generated captions, and the sentence-level metric CHAIR_s reflects the proportion of generated captions that contain hallucinations. In addition, we assess the semantic completeness of generated captions with the metric Recall. Their formulas are defined as follows:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \tag{12}$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinations}\}|}{|\{\text{all sentences}\}|}. \tag{13}$$

$$\text{Recall} = \frac{|\{\text{accurately mentioned objects}\}|}{|\{\text{ground-truth objects}\}|}. \tag{14}$$

Metrics on AMBER. For generative tasks in the AMBER benchmark, the official evaluation metrics include CHAIR (equivalent to CHAIR_i in Eq. 12, hence we use CHAIR_i in the main text), Cover, Hal, and Cog. Let $R_{\text{obj}} = \{r_1, r_2, \ldots, r_m\}$ denote the set of objects mentioned in the response generated by the LVLM, and $G_{\text{obj}} = \{g_1, g_2, \ldots, g_n\}$ denote the set of ground truth objects annotated as present in the image. The specific calculation methods for each metric are as follows:

(1) **CHAIR** evaluates the frequency of hallucinated objects out of all mentioned objects in the model's response, and is defined as:

$$\text{CHAIR} = 1 - \frac{|R_{\text{obj}} \cap G_{\text{obj}}|}{|R_{\text{obj}}|}. \tag{15}$$

(2) **Cover** assesses the degree to which ground-truth objects are accurately captured within the model's generated response. Specifically, it quantifies the proportion of ground-truth objects that are faithfully represented in the output. The formula for Cover is defined as follows:

$$\text{Cover} = \frac{|R_{\text{obj}} \cap G_{\text{obj}}|}{|G_{\text{obj}}|}. \tag{16}$$

(3) **Hal** quantifies the proportion of hallucinated responses relative to the total responses generated by the model, computed as follows:

$$\text{Hal} = \begin{cases} 1, & \text{if CHAIR} \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{17}$$

(4) **Cog** assesses the similarity between the hallucinations generated by LVLMs and those commonly observed in human responses. To this end, the authors annotated a set of objects $H_{\text{obj}} = \{h_1, h_2, \ldots, h_p\}$ that humans are prone to hallucinate. Cog is calculated as:

$$\text{Cog} = \frac{|R_{\text{obj}} \cap H_{\text{obj}}|}{|R_{\text{obj}}|}. \tag{18}$$

The final evaluation scores are obtained by computing the average values of these four metrics across all queries in the AMBER benchmark.

In our experiments, we employ the following versions of the libraries and models: NLTK version 3.9.1, SpaCy version 2.3.9, and the large English language model en_core_web_lg, version 2.3.0.

B.2 Licensing

Benchmarks license. POPE (Li et al., 2023) is licensed under the MIT License. MME (Fu et al., 2023) is restricted for academic research purposes only. CHAIR (Rohrbach et al., 2018) is distributed under the BSD 2-Clause License. AMBER (Wang et al., 2023a) is released under the CC0 license.

LVLMs license. LLaVA-1.5 (Liu et al., 2024b) and LLaVA-NEXT (Liu et al., 2024a) are licensed under the Apache-2.0 License. Qwen-VL (Bai et al., 2023b) permits academic use.

Baselines license. VCD (Leng et al., 2024) is licensed under the Apache-2.0 License. M3ID (Favero et al., 2024) is available for academic use. AvisC (Woo et al., 2024) is licensed under MIT License.

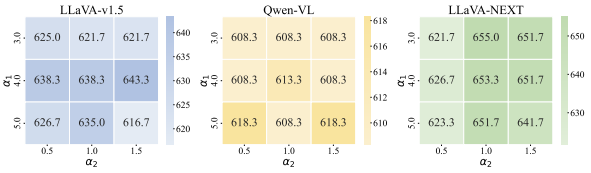


Figure 5: **Ablation study of the decoding coefficient α_1 and α_2** on MME benchmark. We set $\alpha_1 = 4.0$ and $\alpha_2 = 1.0$ for all three evaluated LVLMs. For reference, the optimal performances achieved by the baseline methods in our experiments were 596.7, 593.3, and 613.3 for each respective model.

B.3 Results

Detailed results on POPE. We conduct a comprehensive evaluation of MoD on the POPE benchmark across three datasets: MS-COCO, A-OKVQA, and GQA. POPE consists of 500 images from each dataset, with each image associated with six probing questions, resulting in a total of 27,000 query-answer pairs. As detailed in Tab. 6, MoD demonstrates consistent and superior performance compared to other decoding methods across all 27 experimental settings. Notably, MoD achieves substantial improvements in both Accuracy and F1 score compared to the second-best methods, with maximum performance gains of 3.7 and 4.1 points for LLaVA-1.5, 3.0 and 1.9 points for Qwen-VL, and 1.3 and 2.0 points for LLaVA-NEXT, highlighting its effectiveness in reducing hallucinations and generating reliable outputs.

B.4 Ablation Studies

As previously discussed, MoD distinguishes itself from other decoding strategies by employing a shared set of hyperparameters across diverse tasks and model architectures. Furthermore, MoD exhibits low sensitivity to variations in hyperparameters, as detailed in the following analysis.

Analysis of the decoding coefficient α_1 and α_2. In Fig. 5, we present the results of the ablation study on the impact of the complementary decoding coefficient α_1 and the contrastive decoding coefficient α_2 in Eq. 6 on MME benchmark. The total MME Score is evaluated by simultaneously varying α_1 within the range of 3.0 to 5.0 and α_2 within the range of 0.5 to 1.5, and we observe that MoD exhibits highly stable performance across the tested parameter configurations. Specifically, the maximum deviations are 3.4% for LLaVA-1.5, 0.8% for Qwen-VL, and 4.8% for LLaVA-NEXT, compared to the results obtained using our chosen configura-

Dataset	Setting	Method	LLaVA-v1.5				Qwen-VL				LLaVA-NEXT			
			Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
MS-COCO	random	sampling	83.8	82.4	86.1	84.2	84.9	96.0	72.9	82.9	84.4	94.7	72.8	82.3
		VCD	85.0	82.7	86.1	84.2	85.5	96.0	71.1	83.6	86.0	96.5	74.8	84.3
		M3ID	86.1	83.2	86.8	85.0	85.3	95.1	74.2	83.4	85.5	96.3	73.9	83.6
		AvisC	82.3	78.4	89.3	83.5	82.9	96.2	68.5	80.0	85.2	98.6	71.3	82.8
		MoD (Ours)	**89.2**	90.0	88.2	**89.1**	**86.0**	97.0	74.3	**84.1**	**86.6**	97.7	74.9	**84.8**
	popular	sampling	82.0	79.7	85.9	82.6	84.0	94.7	72.1	81.9	83.2	90.9	73.8	81.5
		VCD	82.1	78.5	88.3	83.2	84.9	94.5	74.9	83.6	84.5	92.9	74.8	82.9
		M3ID	82.8	80.1	88.5	84.1	84.2	94.1	73.8	82.7	84.2	93.1	73.9	82.4
		AvisC	78.2	72.7	90.3	80.5	82.8	95.5	68.9	80.1	83.9	94.7	71.7	81.6
		MoD (Ours)	**85.7**	84.1	88.1	**86.1**	**85.6**	96.3	74.0	**83.7**	**85.5**	95.1	74.9	**83.8**
	adversarial	sampling	75.8	71.3	86.3	78.1	82.1	90.0	72.3	80.2	79.5	84.1	72.9	78.1
		VCD	76.3	71.5	87.3	78.7	84.0	90.6	74.9	82.0	80.9	85.2	74.8	79.7
		M3ID	77.1	71.8	87.6	78.9	83.2	90.4	73.1	80.8	80.6	85.4	73.9	79.2
		AvisC	74.2	68.4	89.9	77.7	81.2	91.9	68.5	78.5	81.8	91.3	71.5	80.2
		MoD (Ours)	**79.7**	75.4	88.2	**81.3**	**84.0**	92.4	74.2	**82.3**	**82.4**	88.3	74.8	**81.0**
A-OKVQA	random	sampling	81.8	76.4	92.1	83.5	86.8	93.2	79.5	85.8	83.8	87.2	79.2	83.0
		VCD	81.2	75.2	93.0	83.2	87.4	92.9	81.1	86.6	84.8	89.2	79.3	83.9
		M3ID	82.9	76.8	94.1	84.6	87.1	92.4	80.3	85.9	85.3	90.7	78.7	84.3
		AvisC	79.1	71.9	95.5	82.1	84.7	93.0	74.9	83.0	85.8	93.4	76.9	84.4
		MoD (Ours)	**86.5**	81.3	94.7	**87.5**	**87.8**	94.4	80.4	**86.9**	**86.4**	91.0	80.7	**85.6**
	popular	sampling	75.3	69.1	91.5	78.7	85.6	90.6	79.5	84.7	81.4	83.4	78.3	80.8
		VCD	74.7	68.2	92.5	78.5	86.3	89.5	81.2	85.1	81.5	82.6	79.9	81.2
		M3ID	75.8	69.8	92.1	79.4	85.9	90.7	79.2	84.6	82.2	84.7	78.7	81.6
		AvisC	71.8	64.7	95.6	77.2	83.9	90.9	75.5	82.4	83.9	90.1	76.5	82.8
		MoD (Ours)	**79.5**	72.6	94.5	**82.2**	**86.5**	91.6	80.3	**85.6**	**84.2**	86.7	80.7	**83.6**
	adversarial	sampling	67.4	61.8	91.2	73.7	80.4	80.1	80.9	80.5	73.2	71.0	78.4	74.5
		VCD	68.1	61.9	93.8	74.6	80.7	80.1	81.6	80.8	74.7	72.2	80.3	76.0
		M3ID	68.3	62.1	93.4	74.6	80.5	80.7	80.2	80.4	74.6	72.5	79.3	75.7
		AvisC	64.4	58.8	96.1	73.0	78.1	80.1	74.7	77.3	75.6	76.5	76.7	76.6
		MoD (Ours)	**69.1**	62.6	94.7	**75.4**	**81.0**	81.2	80.6	**80.9**	**75.7**	73.3	80.9	**76.9**
GQA	random	sampling	81.6	75.6	93.2	83.5	81.3	88.8	71.5	79.2	83.1	85.8	79.4	82.5
		VCD	82.2	76.0	94.1	84.1	82.0	87.6	74.5	80.5	84.0	87.0	80.1	83.4
		M3ID	83.3	76.8	94.0	84.5	82.4	88.1	72.8	79.7	84.3	88.2	79.3	83.5
		AvisC	79.0	71.4	96.7	82.2	80.5	89.9	68.6	77.8	84.9	92.7	75.9	83.4
		MoD (Ours)	**86.2**	80.8	95.1	**87.4**	**83.8**	90.3	75.7	**82.3**	**86.2**	90.3	81.2	**85.5**
	popular	sampling	73.1	66.7	92.5	77.5	75.9	78.1	72.0	74.9	78.5	78.7	78.2	78.5
		VCD	71.5	64.7	94.5	76.8	75.9	76.6	74.7	75.6	78.2	77.2	80.1	78.6
		M3ID	72.3	64.9	94.8	77.1	76.8	78.9	75.2	77.0	78.8	78.5	79.3	78.9
		AvisC	67.4	60.9	97.1	74.8	74.2	77.9	67.5	72.3	80.5	83.8	75.6	79.5
		MoD (Ours)	**74.0**	66.8	95.3	**78.6**	**79.8**	82.4	75.7	**78.9**	**81.1**	81.0	81.2	**81.1**
	adversarial	sampling	68.0	62.0	93.4	74.5	75.5	77.8	71.2	74.4	73.3	71.3	78.0	74.5
		VCD	67.6	61.5	94.4	74.5	76.7	77.8	74.7	76.2	74.2	71.8	79.8	75.6
		M3ID	67.2	61.0	93.9	74.0	77.1	78.6	74.9	76.7	74.2	72.1	79.1	75.4
		AvisC	64.1	58.5	96.7	72.9	75.5	80.5	67.4	73.4	75.8	76.2	75.6	75.9
		MoD (Ours)	**68.7**	62.2	95.3	**75.3**	**78.9**	81.0	75.4	**78.1**	**76.0**	73.6	81.2	**77.2**

Table 6: **Detailed evaluation results on POPE benchmark.** "Acc", "Pre", "Rec", and "F1" stand for Accuracy, Precision, Recall, and F1 score, respectively.

Model	β	Object-level		Attribute-level		Total
		Existence	Count	Position	Color	
LLaVA-v1.5	0.3	195.0	146.7	131.7	160.0	633.3
	0.4	195.0	151.7	131.7	155.0	633.3
	0.5*	195.0	141.7	126.7	175.0	**638.3**
	0.6	195.0	145.0	138.3	155.0	633.3
	0.7	195.0	143.3	133.3	155.0	626.7
Qwen-VL	0.3	160.0	160.0	108.3	180.0	608.3
	0.4	165.0	160.0	108.3	185.0	**618.3**
	0.5*	170.0	160.0	103.3	180.0	613.3
	0.6	170.0	155.0	103.3	180.0	608.3
	0.7	165.0	155.0	103.3	185.0	608.3
LLaVA-NEXT	0.3	195.0	146.7	126.7	165.0	633.3
	0.4	195.0	155.0	121.7	170.0	641.7
	0.5*	195.0	160.0	133.3	165.0	**653.3**
	0.6	195.0	146.7	143.3	160.0	645.0
	0.7	195.0	160.0	138.3	155.0	648.3

Table 7: **Ablation study of the adaptive plausible constraints coefficient** β on MME benchmark. For reference, the best total MME Score achieved by the baseline methods in our experiments were 596.7, 593.3, and 613.3 for each respective model. *We choose $\beta = 0.5$ for all three evaluated LVLMs.

Model	λ	Object-level		Attribute-level		Total
		Existence	Count	Position	Color	
LLaVA-v1.5	0.10	195.0	138.3	131.7	150.0	615.0
	0.15	195.0	140.0	126.7	155.0	616.7
	0.20*	195.0	141.7	126.7	175.0	638.3
	0.25	195.0	160.0	118.3	160.0	**633.3**
	0.30	190.0	163.3	128.3	150.0	631.7
Qwen-VL	0.10	170.0	155.0	103.3	175.0	603.3
	0.15	165.0	160.0	103.3	175.0	603.3
	0.20*	170.0	160.0	103.3	180.0	613.3
	0.25	170.0	165.0	103.3	180.0	**618.3**
	0.30	165.0	150.0	108.3	180.0	603.3
LLaVA-NEXT	0.10	190.0	155.0	93.3	170.0	608.3
	0.15	195.0	153.3	120.0	170.0	638.3
	0.20*	195.0	160.0	133.3	165.0	**653.3**
	0.25	195.0	138.3	116.7	180.0	630.0
	0.30	195.0	146.7	128.3	175.0	645.0

Table 8: **Ablation study of the proportion** λ **for selecting attended image tokens** on MME benchmark. For reference, the best total MME Score achieved by the baseline methods in our experiments were 596.7, 593.3, and 613.3 for each respective model. *We choose $\lambda = 0.20$ for all three evaluated LVLMs.

tion, indicating the robustness of MoD with respect to variations in α_1 and α_2.

Analysis of the adaptive plausible constraints coefficient β. In Tab. 7, we conduct an ablation study to investigate the impact of the adaptive plausible constraints coefficient β in Eq. 8 on the MME benchmark. By varying β within the range of 0.3 to 0.7, we observe that the total MME Score exhibits remarkable stability across different models. Specifically, the maximum deviations are only 1.8% for LLaVA-1.5, 0.8% for Qwen-VL, and 3.1% for LLaVA-NEXT when compared to the results obtained using our selected β value. Notably, even the least favorable results still surpass the best baseline methods, underscoring the robustness of our approach concerning the choice of β.

Analysis of the proportion λ for selecting attended image tokens. In Tab. 8, we conduct an ablation study to evaluate the influence of the proportion λ for selecting attended image tokens defined in Eq. 3. Experimental results demonstrate that MoD achieves remarkable stability across a wide range of λ values, from 0.10 to 0.30. Notably, we observe a more pronounced degradation in performance when λ is decreased compared to when it is increased from 0.20. This asymmetry can be attributed to the fact that decreasing λ leads to a diminished amount of visual information for both complementary and contrastive strategies, thereby affecting the overall performance.

B.5 Comparison with DeGF

As an adaptive decoding method, we compared MoD with DeGF (Zhang et al., 2025) on MME. DeGF first generates an image based on the original output, then adaptively selects a contrastive decoding strategy by evaluating the consistency between the generated image's output and the original output. In contrast, MoD adaptively chooses its contrastive decoding strategy based on the consistency between the original output and the output derived from only partial image tokens. Experimental results show that DeGF slightly outperforms MoD overall, with each method having its own advantages in different subcategories. However, DeGF requires an additional inference step and an image generation process compared to MoD. Therefore, in practice, the choice of decoding strategy should balance computational cost and performance based on specific needs.

Method	Object-level		Attribute-level		Total
	Existence	Count	Position	Color	
sampling	170.0	103.3	108.3	128.3	510.0
DeGF	190.0	150.0	133.3	170.0	643.3
MoD	195.0	141.7	126.7	175.0	638.3

Table 9: **Comparison with DeGF using LLaVA-v1.5.**

C More Case Studies

We provide additional case studies of generative tasks involving all three evaluated LVLMs, which generate responses of varying lengths. As shown

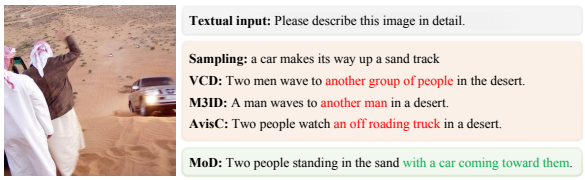


Figure 6: **Case study of generative tasks using Qwen-VL.** We compare responses generated by sampling, VCD, M3ID, AvisC, and our proposed MoD. Hallucinated content is highlighted in red, while more detailed and accurate content is marked in green.

in Fig. 6, MoD accurately identifies that the car is coming toward the two people. In Fig. 7, other decoding methods exhibit severe hallucinations, such as referring to "a few sneakers" or "a backpack nearby". Conversely, MoD effectively mitigates these hallucinations and provides an exact description of the girls' location and orientation. In Fig.8, baseline methods tend to generate speculative content regarding the computer screen display, despite inadequate image resolution. These methods are also prone to hallucinate objects that commonly co-occur with computers, like "coffee mug", "printer" and so on. In contrast, MoD accurately generates an overview of the screen instead, without introducing any hallucinations, demonstrating its superior ability to generate accurate and detailed captions.

Textual input: Please describe this image in detail.

Sampling: The image shows a view of a beach, capturing the essence of an ocean shoreline. A young person wearing a pink coat is standing near the water, possibly tying their shoes, while two children on the beach appear to be playing. One child is near the center of the scene, and the other is slightly to the left. In the scene, there are a few sneakers visible on the sand, with one pair placed closer to the left side of the beach, another pair near the center, and the third pair slightly to the right. The connection of the children with the beach and the ocean creates a vibrant, joyful atmosphere.

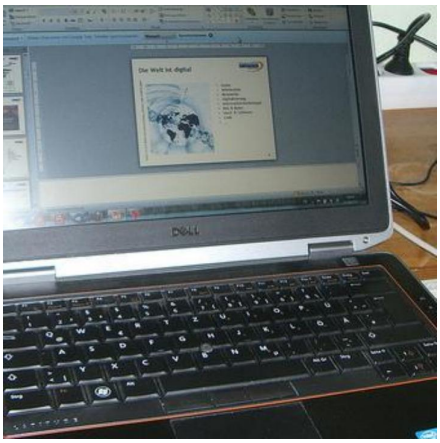
VCD: The image shows two children, a boy and a girl, standing on the sandy beach near the ocean. They are both wearing pink jackets and green boots, and they appear to be playing or exploring the area. The children are looking towards the ocean, enjoying their time together. The beach scene features a few sand footprints and a backpack nearby, possibly belonging to the children or other beachgoers. The sky can be seen above, providing a beautiful backdrop for the children's playtime by the water.

M3ID: The image depicts a beautiful beach scene with two little girls standing on the sand, one closer to the water and the other further away. Both are wearing jackets, and one of them is bending down on the sand. They seem to be playing near the ocean or possibly preparing to enter the water. The ocean is visible in the background, with a cloudy sky casting its shadow over the beach. The girls' footprints can be seen on the sand, creating a unique atmosphere on the beach.

AvisC: In the image, there are two young girls enjoying their time on a sandy beach near the ocean. One girl is standing closer to the water, while the other is located further back on the sand. Both girls are looking out at the waves, likely admiring the ocean view. There are several backpacks scattered around the scene, with two close to the first girl and another near the second girl. Additionally, a surfboard can be seen resting on the sand, suggesting that the girls might be interested in surfing or other water-based activities.

MoD: The image captures a beautiful beach scene with two children playing near the ocean. One child is standing closer to the water, while the other is positioned further back on the sandy shore. They seem to be enjoying their time together, possibly playing in the sand or observing the waves. In the background, the ocean can be seen, with waves gently rolling towards the shore. The beach setting and the presence of the children create a peaceful and joyful atmosphere.

Figure 7: **Case study of generative tasks using LLaVA-1.5.** We compare responses generated by sampling, VCD, M3ID, AvisC, and our proposed MoD. Hallucinated content is highlighted in red, while more detailed and accurate content is marked in green.



Textual input: Please describe this image in detail.

Sampling: The image depicts an open, black Dell laptop on a wooden desk, powered on and displaying a presentation with text and images related to automated dialing. There are two slides visible on the screen: the left displays a colorful logo with the text "Telemedix" and a graphic of a human profile with medical equipment, while the right slide has a white background with the title "Automated Dialing for Medical Sales" in bold purple letters, followed by bullet points outlining its benefits and features such as "Efficient Call Generation", "Accurate Data Collection", "Easy System Integration", and "Reduced Administrative Tasks". The laptop is connected to a black USB cable that extends into the upper left corner of the shot. To the right of the laptop, a red laptop stand and a single black speaker are also visible. The overall setting suggests a workspace or conference room environment. The image is a color photograph with a focus on the equipment for illustrative purposes.

VCD: The image captures a moment of productivity, featuring a Dell laptop on a wooden desk. The laptop screen displays a presentation slide titled "Digitization: A Work in Progress", indicating that the user might be engaged in a workshop, conference, or seminar related to digital transformation. The slide itself is white with black text and graphics, providing a clear contrast against the laptop's screen. A coffee mug, presumably filled with a warm beverage, sits to the right of the laptop, suggesting that the user may be enjoying a drink while working. The mug is brown, matching the color of the wooden desk. In the background, a white power strip and a white phone cord can be seen, indicating that the user might be using multiple devices. The overall setting suggests a comfortable and organized workspace, conducive to focus and productivity.

M3ID: The image shows a laptop sitting on a wooden desk with a power cord plugged into a nearby power outlet. The laptop screen displays a presentation slide with text and graphics related to oceanography and marine technology. The slide is titled "Digital Workshop" and appears to be from a presentation, with a company logo visible. There's a coffee mug to the right of the laptop. The laptop brand is Dell, and it features an orange backlit keyboard with a touchpad. The laptop has a slim profile with a black front cover and silver edges, giving it a sleek, modern design. There's no visible user activity on the screen. The desk surface is clean, and the surrounding environment seems to be a home or office setting. The photo captures the laptop at an angle where the screen is clearly visible and the presentation is displayed on the laptop's monitor. The style of the image is straightforward, documenting an everyday scenario of technology usage.

AvisC: In the heart of a cozy room, a black Dell laptop sits open on a wooden desk, its screen alive with the glow of an Excel spreadsheet. The spreadsheet, bathed in white, hosts a presentation titled "Water Quality Monitoring", complete with a vibrant blue bar graph that illustrates data points. Attached to the laptop via a power cord, a black printer waits patiently for its next task. It shares the space with a white mug, perhaps filled with a warm beverage to fuel the user's work. Behind this technological trio, a window draped with white curtains lets in soft light, illuminating the scene and casting a peaceful atmosphere over the workspace. The relative positions of these objects suggest a well-organized and functional workspace, ready for any challenge that comes its way.

MoD: The image shows a laptop computer with its screen open, displaying a presentation slide with a title and bullet points. The laptop is placed on a wooden surface, possibly a table or desk. The screen of the laptop is turned on, and the slide appears to be a PowerPoint presentation. The laptop is a Dell model, as indicated by the logo below the screen. The background is not detailed, but it seems to be an indoor setting with a wooden surface, suggesting that someone might be working or studying in this area. The overall impression is of a workspace or study area.

Figure 8: **Case study of generative tasks using LLaVA-NEXT.** We compare responses generated by sampling, VCD, M3ID, AvisC, and our proposed MoD. Hallucinated content is highlighted in red, while more detailed and accurate content is marked in green.