MoLE: Decoding by Mixture of Layer Experts Alleviates Hallucination in Large Vision-Language Models

Tian Liang[1], Yuetian Du[1], Jing Huang[1], Ming Kong[1], Luyuan Chen[2], Yadong Li[3], Siye Chen[3], Qiang Zhu[1]*

[1]Zhejiang University,
[2]Beijing Information Science and Technology University,
[3]Ant Group
{liangtian2022, 22421227, huangjin9, zjukongming, zhuq}@zju.edu.cn; chenly@bistu.edu.cn; {liyadong.lyd, chensiye.csy}@antgroup.com

Abstract

Recent advancements in Large Vision-Language Models (LVLMs) highlight their ability to integrate and process multi-modal information. However, hallucinations—where generated content is inconsistent with input vision and instructions—remain a challenge. In this paper, we analyze LVLMs' layer-wise decoding and identify that hallucinations can arise during the reasoning and factual information injection process. Additionally, as the number of generated tokens increases, the forgetting of the original prompt may also lead to hallucinations.To address this, we propose a training-free decoding method called Mixture of Layer Experts (MoLE). MoLE leverages a heuristic gating mechanism to dynamically select multiple layers of LVLMs as expert layers: the Final Expert, the Second Opinion expert, and the Prompt Retention Expert. By the cooperation of each expert, MoLE enhances the robustness and faithfulness of the generation process. Our extensive experiments demonstrate that MoLE significantly reduces hallucinations, outperforming the current state-of-the-art decoding techniques across three mainstream LVLMs and two established hallucination benchmarks. Moreover, our method reveals the potential of LVLMs to independently produce more reliable and accurate outputs.

Code — https://github.com/Rainlt/MoLE/

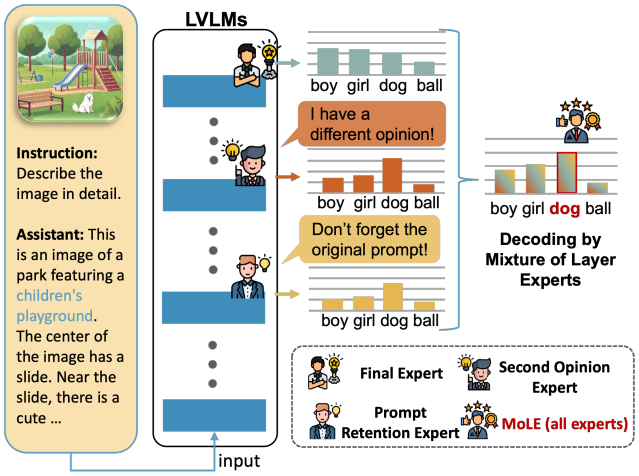

Figure 1: An illustration of **Mixture of Layer Experts(MoLE)** Decoding. We designed heuristic gating mechanisms to select three expert layers from LVLMs: the Final Expert, Second Opinion Expert, and Prompt Retention Expert. By employing collaborative decoding from multiple experts, we achieve more accurate and faithful answers, as illustrated in the example of 'dog' in the figure.

Introduction

Large Vision-Language Models (LVLMs) have emerged as the dominant paradigm for tackling vision-language tasks, demonstrating remarkable performance across various applications. However, a critical challenge that continues to hinder the effectiveness of these models is the issue of hallucination—where the generated content deviates from the input vision or instructions. This not only compromises the faithfulness of the output but also undermines the reliability of LVLMs in real-world scenarios (Liu et al. 2024b). Addressing this problem has become a pressing concern in the field.

Existing approaches to mitigating hallucinations have primarily focused on contrastive decoding techniques (Li et al. 2023b), which leverage "amateur models" for eliminating erroneous outputs. These methods have gained traction due to their ability to reduce hallucinations without the need for additional training. Typically, in such setups, an expert model (the original LVLM) is contrasted with a weaker or more confused model, and the differences in their outputs are used to filter out hallucinations. While this approach has shown some success, it is inherently limited by the reliance on weaker models, which may not always provide the most accurate guidance.

Recognizing the limitations of traditional contrastive decoding, we draw inspiration from the Mixture of Experts (MoE) (Jacobs et al. 1991) framework to propose a novel strategy, **Mixture of Layer Experts (MoLE) Decoding**, aimed at addressing the hallucination problem more effectively. Unlike previous methods that construct amateur models, our approach leverages multiple complementary expert

*Corresponding author.

models, each specialized in different aspects of the decoding process. By introducing carefully designed gating mechanisms, we enable these experts to collaboratively generate outputs that are more faithful to the input vision and instructions.

While employing multiple expert models typically introduces additional computational costs, we mitigate this by utilizing the existing LVLM structure to obtain multiple collaborative experts within a single forward pass, thus maintaining computational efficiency while enhancing output faithfulness.

Our approach is informed by insights from recent studies on model interpretability and layer-wise information processing. Prior research has shown that language models encode low-level features in shallow layers and more abstract, semantic information in deeper layers (Tenney, Das, and Pavlick 2019). Additionally, findings from model editing suggest that factual information, or "world knowledge", is embedded in the mid-to-late layers of models (Zheng et al. 2023a). Building on these insights, we hypothesized that different layers of an LVLM could serve as distinct experts, each contributing uniquely to the decoding process.

Consequently, we represents the **Mixture of Layer Experts (MoLE)**, a method specifically designed to mitigate hallucinations by leveraging expert mixtures drawn from different layers of an LVLM. As illustrated in Figure 1, our method identifies three key experts: the *Final Expert* from the last layer, responsible for refining the final output; the *Second Opinion (SO) Expert*, selected from the final layers to provide alternative insights for reference; and the *Prompt Retention (PR) Expert*, drawn from layers that best retain the original prompts, ensuring that the model's output remains true to the input vision and instructions. By orchestrating these experts through a single forward pass, MoLE effectively reduces hallucinations with minimal computational overhead.

In summary, our contributions are as follows:

- We introduce the Mixture of Experts (MoE) (Fedus, Zoph, and Shazeer 2022) concept into the LVLM decoding process, shifting from traditional amateur model-based contrastive decoding to a collaborative, layer-wise approach that significantly enhances output faithfulness.
- We develop a novel layer-wise expert decoding method, **MoLE**, which leverages the Final Expert, Second Opinion Expert, and Prompt Retention Expert, each derived from different layers of an LVLM, combined with tailored gating mechanisms for improved decoding accuracy.
- Through extensive experiments on three leading LVLMs across two multimodal hallucination benchmarks, we validate the effectiveness of MoLE, showing it significantly reduces hallucinations while maintaining computational efficiency without relying on additional tools.

Related Work

Large Vision-Language Models

Large Vision-Language Models (LVLMs) have made significant strides in integrating visual and linguistic information, thereby enabling a unified approach to tasks such as image captioning, visual question answering, and image-text retrieval. Leveraging the power of transformers (Vaswani et al. 2017), foundational models like BERT (Devlin et al. 2019) and large language models (LLMs) such as LLaMA (Touvron et al. 2023a,b) and Vicuna (Zheng et al. 2023b) have laid the groundwork for deeper vision-language integration. Models like CLIP (Radford et al. 2021) and query-based methods like BLIP (Li et al. 2022, 2023a; Dai et al. 2023; Liang et al. 2024) have effectively aligned text and image features, paving the way for the development of robust LVLMs capable of handling complex multimodal tasks.

Recent advances in LVLMs, including MiniGPT-4 (Zhu et al. 2023), LLaVA-1.5 (Liu et al. 2023), and Shikra (Chen et al. 2023), have further scaled model sizes and capabilities, enhancing their ability to process and generate content across multiple modalities. However, despite these advancements, a persistent challenge remains: hallucinations, where models generate content not grounded in the input data. Addressing this issue is critical to improving the reliability and faithfulness of LVLM outputs, particularly in real-world applications.

Hallucination in LVLMs

Hallucination in LVLMs, originally identified in natural language processing, refers to the generation of content that either deviates from the provided context (faithful hallucinations) or is factually incorrect (factual hallucinations) (Huang et al. 2023). In the context of LVLMs, object hallucinations—where models incorrectly identify or describe objects not present in the visual input—are particularly problematic (Rawte, Sheth, and Das 2023; Liu et al. 2024c).

Efforts to mitigate hallucinations have focused on various strategies, including dataset enhancements (Liu et al. 2024a), architectural modifications (Liu et al. 2023; Li et al. 2024), inference techniques (et al 2023; Chuang et al. 2023; Huang et al. 2024; Zhao et al. 2024; Favero et al. 2024), and post-processing methods (Zhou et al. 2023; Yin et al. 2023). Among these, inference-phase techniques have gained prominence due to their lightweight and effective nature, often eliminating the need for additional training. Contrastive decoding (Li et al. 2023b), a key strategy in this domain, contrasts outputs from an "amateur" model with those from the original "expert" model to filter out hallucinations. While methods like VCD (et al 2023) and M3ID (Favero et al. 2024) utilize noise or omission strategies to create amateur models, they often come at the cost of increased inference overhead. DoLA (Chuang et al. 2023) designed for LLM hallucinations attempts to reduce this by using shallow outputs from models as amateur models, yet it falls short in addressing faithful hallucinations in LVLMs.

In response to the limitations, we propose a novel decoding method, Mixture of Layer Experts (MoLE). Unlike existing approaches that rely on amateur models, MoLE leverages multiple expert layers within LVLMs to collaboratively enhance output faithfulness. This method not only reduces hallucinations but does so with minimal computational overhead, as demonstrated through extensive experimentation.

Method

Overview

In this section, we introduce the Mixture of Layer Experts (MoLE) decoding framework designed to mitigate hallucinations in Large Vision-Language Models (LVLMs). Our approach leverages multiple expert layers within the LVLM to enhance the faithfulness of generated outputs. We begin by detailing the decoding process of a typical LVLM, followed by a description of our method to select and combine expert layers: the Final Expert, the Second Opinion Expert, and the Prompt Retention Expert.

LVLM Decoding and Final Expert

A typical LVLM consists of an embedding layer, a stack of N transformer blocks, and a final classification head ϕ. The classification head is used to predict the logits distribution over the vocabulary set χ for the next token. When generating a sequence, at each time step t, the LVLM takes in the prompt $PT = \{pt_1, \ldots, pt_m\}$ which consists of vision features and text instructions, and the previous tokens $\{x_1, x_2, \ldots, x_{t-1}\}$ generated before time step t. These inputs are embedded into a sequence $H_0 = \{hp_1^{(0)}, \ldots, hp_m^{(0)}; h_1^{(0)}, \ldots, h_{t-1}^{(0)}\}$ by the embedding layer.

The sequence H_0 then passes through N transformer blocks, producing outputs at each layer, denoted as H_j for the j-th layer. The output of the final layer $H_N = \{hp_1^{(N)}, \ldots, hp_m^{(N)}; h_1^{(N)}, \ldots, h_{t-1}^{(N)}\}$ is passed to the final classification head ϕ, which generates the logits distribution for the next token x_t:

$$q_N(\cdot \mid PT; x_{<t}) = (\phi(h_{t-1}^{(N)}))_{x_t}, \quad x_t \in \chi \tag{1}$$

The probability distribution is obtained by applying softmax to the logits:

$$p(x_t \mid PT; x_{<t}) = \text{SoftMax}(q_N(\cdot \mid PT; x_{<t})) \tag{2}$$

The final layer N is typically used as the Final Expert due to its capacity to synthesize information across all preceding layers. However, relying solely on this expert can lead to hallucinations, especially in complex multimodal tasks. Therefore, our method introduces additional expert layers to assist in generating more accurate and contextually faithful outputs.

Second Opinion Expert

In many decision-making processes, particularly in fields like medicine and law, a second opinion is often sought to reduce errors and provide a more balanced perspective. Analogously, we introduce a Second Opinion Expert within the LVLM framework. This expert is selected from one of the last $\mathcal{L}$ layers, which are known to encode different levels of world knowledge (Zheng et al. 2023a), accumulated during the pre-training process. As shown in Figure 2, the final expert introduces the hallucination, along with the upheaval of JSD, indicating the injection of world knowledge.

The key idea behind the Second Opinion Expert is to introduce a layer that can provide a different perspective from the Final Expert, particularly on critical tokens, while maintaining consistency on the majority of tokens. This is achieved by evaluating the divergence between the logits of the final layer and those of the candidate layers. We use Jensen-Shannon Divergence (JSD) to measure the difference between these logits distributions:

$$\begin{aligned} &d\left(q_N(\cdot \mid PT, x_{<t}), q_j(\cdot \mid PT, x_{<t})\right) \\ &\quad = \text{JSD}\left(q_N(\cdot \mid PT, x_{<t}) \parallel q_j(\cdot \mid PT, x_{<t})\right) \end{aligned} \tag{3}$$

Here, $q_N(\cdot \mid PT, x_{<t})$ represents the logits output from the final layer N, and $q_j(\cdot \mid PT, x_{<t})$ represents the logits from the j-th layer.

Gating For Second Opinion Expert: To ensure that the SO Expert provides valuable opinion, we select a layer that shows maximum divergence on critical (Top-k) tokens, while being consistent with the Final Expert on the majority of tokens. To simplify the formula, we denote the logits corresponding to the top-k tokens predicted by the j-th layer $q_j\left(x_{top-k} \mid PT, x_{<t}\right)$ as $q_j^{(top-k)}$, and the logits outside the top-k predictions $q_{\text{j}}\left(x_{\text{majority}} \mid PT, x_{<t}\right)$ as $q_j^{(majority)}$. Hence the formula:

$$M_{\text{top-k}} = \arg\max_{j \in \mathcal{N}} d\left(q_N^{(top-k)}, q_j^{(top-k)}\right) \tag{4}$$

$$M_{\text{majority}} = \arg\min_{j \in \mathcal{N}} d\left(q_N^{(majority)}, q_j^{(majority)}\right) \tag{5}$$

$$M_{SO} = \begin{cases} M_{\text{top-k}}, & \text{if } M_{\text{top-k}} = M_{\text{majority}} \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

Where the $\mathcal{N} \in \{0, 1, \ldots, 31\}$ represents the index set of the LVLM. The selected Second Opinion Expert layer, denoted as M_{SO}, is selected based on the criterion that it maximizes the divergence for critical tokens while minimizing it for the majority, as shown in Figure 4(a) Thus the logits generated by this expert can be formulated as:

$$q_{SO} = \alpha \mathbf{1}_{\{M_{top-k} = M_{majority}\}} q_t^{M_{SO}} \tag{7}$$

Here, α is a scale factor controlling the intensity of the expert layer M_{SO}. And $\mathbf{1}_{\{\ldots\}}$ represents the indicator function, which equals 1 when the condition inside the brackets is true, and 0 otherwise. $q_t^{M_{SO}}$ represents the logits from M_{SO} at time step t.

Prompt Retention Expert

As the sequence generation progresses, the model's attention to the initial prompt tends to wane, which can result in hallucinations, especially in longer sequences. This occurs because the model gradually loses the original context provided by the prompt as more tokens are generated. Figure 3 illustrates this: as sequence length increases, attention to the prompt decreases, and the occurrence of hallucinations within each unit length rises, supporting our hypothesis.

To address this, we introduce a Prompt Retention Expert, which is specifically selected based on its ability to maintain high attention to the prompt throughout the sequence generation.

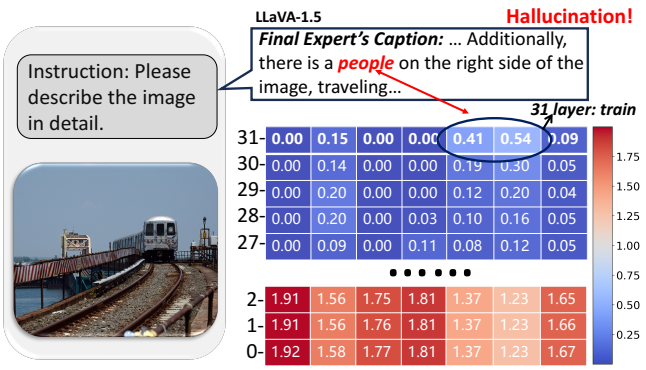

Figure 2: JSD between the final expert layer and early layers. The injection of world knowledge into the final layers of JSD causes drastic changes, leading to the hallucination of "people," whereas the 31st layer is capable of producing the faithful output "train".

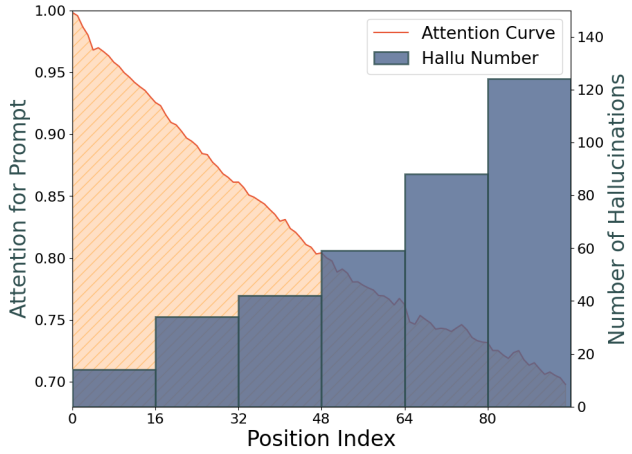

Figure 3: Illustration of the number of hallucinations generated at different positions and the attention decay over the Prompt section.

Gating For Prompt Retention Expert: As shown in Figure 4(b), we calculate the sum of attention scores directed at the prompt tokens in each layer and select the layer with the highest sum as the Prompt Retention Expert:

$$S_t^{(j)} = \sum_{k=1}^{m} \text{softmax}\left(\frac{h_t^{(j)} \cdot (hp_k^{(j)})^\top}{\sqrt{d}}\right) \tag{8}$$

$$M_{PR} = \text{argmax}_{j \in \mathcal{N}} S_t^{(j)} \tag{9}$$

Here, $S_t^{(j)}$ represents the sum of attention scores for the prompt tokens at layer j, and M_{PR} denotes the selected Prompt Retention Expert layer. To ensure that the Prompt Retention Expert's influence grows as the sequence length increases, we apply a time-dependent weight to get the logits of Prompt Retention Expert:

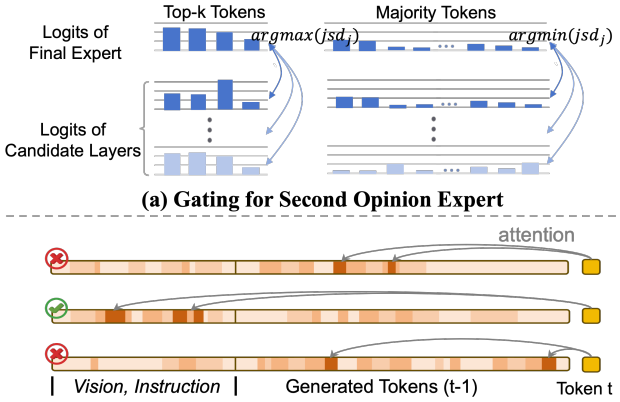

Figure 4: (a) An illustration of gating mechanism of the Second Opinion Expert. jsd_j represents the JSD between the logits of the Final Expert and the j-th layer. (b) An illustration of gating mechanism of the Prompt Retention Expert. The layer with max attention on vision and Instruction will be selected.

$$q_{PR} = \left(1 - e^{-\frac{1}{\lambda}t}\right) \cdot q_t^{PR} \tag{10}$$

Here, q_t^{PR} represents the logits corresponding to M_{PR}, and λ is the temperature coefficient that controls the rate at which q_{PR} increases over time. This time-dependent weight $(1 - e^{-\frac{t}{\lambda}})$ increases the influence of the Prompt Retention Expert as the sequence progresses, thereby mitigating the tendency of the model to drift away from the original prompt.

Mixture of Layer Experts Decoding

Having identified the Final Expert, Second Opinion Expert, and Prompt Retention Expert, we proceed to combine their logits to generate the final prediction. Unlike traditional contrastive decoding methods, which subtract logits from amateur models, our approach sums the logits from these expert layers, leveraging their complementary strengths:

$$p_{mole} = \text{softmax}(q_F + q_{SO} + q_{PR}) \tag{11}$$

In this framework, the Final Expert serves as the primary source of prediction, drawing on the most comprehensive synthesis of information. The Second Opinion Expert introduces a critical perspective on key tokens, ensuring that the model considers alternative interpretations where necessary. Meanwhile, the Prompt Retention Expert ensures that the model stays true to the original prompt, especially as the sequence lengthens.

This collaborative decoding approach effectively reduces hallucinations by combining the expertise of multiple layers within the LVLM. Moreover, the method is computationally efficient, requiring only a single forward pass, and does not necessitate any additional training or external tools. As a result, MoLE offers a practical and powerful solution for

enhancing the faithfulness of outputs in LVLMs, making it suitable for a wide range of complex multimodal tasks.

Experiments

Experimental Settings

Baselines. To evaluate the effectiveness of our MoLE method, we compare it against several existing state-of-the-art training-free decoding approaches that are commonly used in LVLMs to mitigate hallucinations. The baseline methods include:

- Greedy Search and Beam Search: Traditional decoding strategies that are widely used in sequence generation tasks.
- OPERA (Huang et al. 2024): A state-of-the-art method that uses multiple rollbacks and token aggregation suppression to reduce hallucinations.
- VCD (et al 2023): A technique that introduces noise into images to create amateur models for contrastive decoding.
- DoLA (Chuang et al. 2023): A method that leverages layer-wise contrastive decoding to enhance factuality.

LVLM Backbones. To ensure the generalizability of our MoLE approach, we conducted experiments using three state-of-the-art LVLMs: MiniGPT-4 (Zhu et al. 2023), LLaVA-1.5 (Liu et al. 2023), and Shikra (Chen et al. 2023). Each of these models utilizes Vicuna7b (Zheng et al. 2023b) as the language decoder, which is a popular choice for multimodal tasks. We applied all baseline methods and our proposed MoLE method across these three models to provide a robust comparison of performance across different LVLM architectures.

Implementation Details. In our implementation of MoLE, the Final Expert is selected as the last layer of the model ($N = 32$). The Second Opinion Expert is dynamically chosen from the last three layers ($\mathcal{L} \in \{29, 30, 31\}$), excluding the final layer. These layers were selected based on their strong inference capabilities and distinct world knowledge, which aligns with our design principles for the Second Opinion Expert. We set $k = 5$ to determine the top-k critical tokens and used $\alpha = 0.5$ as the scale factor for the Second Opinion Expert. For the Prompt Retention Expert, the temperature coefficient λ was set to 100. The optimality of these parameters was confirmed through ablation studies, and more detailed settings are provided in the Supplementary Material.

Main Results

POPE. To evaluate hallucination reduction, we employed the Polling-based Object Probing Evaluation (POPE) metric (Li et al. 2023c). POPE focuses on object hallucinations but uses a question-and-answer format, asking questions like "Is there <*object*> in the image?" Besides querying objects present in the image, POPE introduces various non-existent objects as negative samples to assess whether the model can correctly identify specific objects in the image or produce hallucinations. The full POPE test comprises three parts, with each part having a 1:1 ratio of positive to negative samples. The "Random" part selects negative samples randomly from existing objects in the dataset; the "Popular" part assesses the most common objects in the dataset; and the "Adversarial" part selects objects frequently co-occurring with real objects as negative samples to evaluate the model's ability to identify highly relevant objects in the image.

Following the procedure in HALC (Chen et al. 2024), we selected 100 images from the COCO dataset and created 600 samples, comprising equal numbers of positive and negative samples for each part of the test.

The results presented in Table 1 showcase the effectiveness of our proposed **MoLE** method across different decoding strategies on the Polling-based Object Probing Evaluation (POPE) dataset. MoLE consistently outperforms other strategies, demonstrating its robustness in enhancing output accuracy and reducing hallucinations in Large Vision-Language Models (LVLMs).

In the "Random" sampling setting, MoLE achieved the highest accuracy and precision across all models, particularly improving MiniGPT-4's accuracy by 8.7% over Beam Search. This suggests that MoLE's layer-wise approach effectively handles diverse object categories and visual contexts, resulting in more accurate and faithful outputs.

For "Popular" sampling, MoLE continued to excel, reducing the tendency of models to hallucinate frequently occurring objects. For instance, MoLE increased MiniGPT-4's precision by 7.7% over DoLA, highlighting its capability to maintain fidelity in scenarios prone to hallucinations.

In the challenging "Adversarial" sampling, MoLE's resilience was evident, outperforming all other methods with significant gains in accuracy and precision. This demonstrates MoLE's suitability for safety-critical applications, where minimizing errors is crucial.

Overall, MoLE's consistent outperformance across all settings underscores its value in improving the reliability of LVLMs. By leveraging multiple expert layers, MoLE effectively mitigates hallucinations, making it a valuable contribution to advancing vision-language tasks in real-world applications.

CHAIR. The CHAIR (Rohrbach et al. 2019) (Caption Hallucination Assessment with Image Relevance) metric is specifically designed to quantify object hallucinations in image captioning tasks. It measures the extent to which generated descriptions mention objects that are not present in the ground truth. CHAIR consists of two components: $CHAIR_S$ (sentence-level hallucination assessment) and $CHAIR_I$ (image-level hallucination assessment). A lower CHAIR value indicates fewer hallucinations, reflecting better model performance.

For the CHAIR evaluation, we randomly sampled 500 images from the MSCOCO (Lin et al. 2014) validation set and instructed each model to generate detailed descriptions of these images. The evaluation follows the methodology outlined in HALC (Chen et al. 2024).

As demonstrated in Table 2, our MoLE method significantly outperforms previous state-of-the-art approaches across all metrics and models. In particular, the MiniGPT-4 (Zhu et al. 2023) model shows a 21% reduction in $CHAIR_I$ compared to DoLA, highlighting the effectiveness of MoLE in mitigating object hallucinations in long text generation tasks.

Setting	Model	Decoding	Accuracy	Precision	F_1 Score
Random	MiniGPT-4	Beam Search	0.750	0.689	0.784
		OPERA	0.747	0.688	0.781
		VCD	0.662	0.627	0.703
		DoLA	0.753	0.713	0.774
		MoLE(ours)	**0.818**	**0.917**	**0.794**
	LLaVA-1.5	Beam Search	0.858	0.812	0.868
		OPERA	0.802	0.731	0.828
		VCD	0.730	0.661	0.777
		DoLA	0.838	0.783	0.853
		MoLE(ours)	**0.887**	**0.858**	**0.891**
	Shikra	Beam Search	0.783	0.776	0.786
		OPERA	0.785	0.790	0.783
		VCD	0.775	0.772	0.776
		DoLA	0.765	0.787	0.756
		MoLE(ours)	**0.825**	**0.857**	**0.817**
Popular	MiniGPT-4	Beam Search	0.648	0.597	0.721
		OPERA	0.645	0.596	0.718
		VCD	0.618	0.587	0.677
		DoLA	0.718	0.674	**0.750**
		MoLE(ours)	**0.735**	**0.751**	0.726
	LLaVA-1.5	Beam Search	0.783	0.718	0.812
		OPERA	0.735	0.664	0.782
		VCD	0.688	0.628	0.748
		DoLA	0.765	0.697	0.799
		MoLE(ours)	**0.822**	**0.766**	**0.839**
	Shikra	Beam Search	0.778	0.768	0.782
		OPERA	0.778	0.779	0.778
		VCD	0.753	0.739	0.761
		DoLA	0.747	0.757	0.741
		MoLE(ours)	**0.810**	**0.832**	**0.803**
Adversarial	MiniGPT-4	Beam Search	0.647	0.596	**0.720**
		OPERA	0.633	0.587	0.711
		VCD	0.620	0.588	0.678
		DoLA	0.658	0.615	0.712
		MoLE(ours)	**0.723**	**0.733**	0.718
	LLaVA-1.5	Beam Search	0.743	0.676	0.784
		OPERA	0.698	0.631	0.760
		VCD	0.655	0.600	0.730
		DoLA	0.725	0.658	0.773
		MoLE(ours)	**0.775**	**0.711**	**0.805**
	Shikra	Beam Search	0.767	0.752	0.773
		OPERA	0.762	0.754	0.765
		VCD	0.718	0.697	0.733
		DoLA	0.727	0.727	0.727
		MoLE(ours)	**0.798**	**0.811**	**0.794**

Table 1: POPE results with random, popular, and adversarial samplings. The best performances within each setting are bolded.

Method	MiniGPT-4		LLaVA-1.5		Shikra	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Greedy	22.20	7.50	<u>20.80</u>	<u>6.60</u>	21.40	<u>7.21</u>
Beam Search	21.60	<u>7.18</u>	22.20	7.40	20.60	8.00
VCD	23.60	8.67	23.40	7.70	21.40	7.56
OPERA	<u>20.40</u>	7.60	21.40	6.97	<u>20.00</u>	7.77
DoLA	24.20	8.20	21.60	6.92	22.00	8.00
MoLE (Ours)	**19.20**	**7.10**	**18.00**	**6.00**	**18.40**	**6.10**

Table 2: CHAIR hallucination evaluation results on three LVLM models. Denote $CHAIR_S$ as C_S and $CHAIR_I$ as C_I . Smaller values corresponds to less hallucinations.

Setup	SO	SOG	PR	PRG	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$
A	-	-	-	-	22.2	7.4
B	✓	-	-	-	22.2	7.1
C	✓	✓	-	-	20.4	6.83
D	-	-	✓	-	19.4	6.97
E	-	-	✓	✓	18.0	6.54
F	✓	✓	✓	✓	**17.2**	**5.63**

Table 3: Ablation on every module of MoLE. SOG represents the gating mechanism for the SO Expert. If it is empty, it indicates that the logits of candidate layers are averaged directly. And PR Expert's weight is fixed at 1 if PRG is empty.

Ablation Study

Impact of Each Expert Module. To evaluate the individual contributions of each component in the MoLE framework, we conducted ablation studies on the LLaVA-1.5 (Liu et al. 2023) model using the CHAIR metric. Our MoLE framework integrates three expert layers: the Final Expert, Second Opinion Expert (SO), and Prompt Retention Expert (PR). Additionally, we implemented gating mechanisms for the SO and PR experts, termed as **Second Opinion Gating (SOG)** and **Prompt Retention Gating (PRG)**, respectively. The results, detailed in Table 3, show that removing all experts results in performance similar to beam search. However, as we incrementally add the Second Opinion Expert and Prompt Retention Expert, hallucinations are significantly reduced. The gating mechanisms further enhance performance, demonstrating their critical role in the MoLE framework. Specifically, without the Second Opinion Gating, the logits from the Second Opinion Expert are averaged across candidate layers. Similarly, without the Prompt Retention Gating, the influence of the PR Expert remains fixed, leading to suboptimal results.

Effectiveness of the SO Gating Mechanism. To further validate the effectiveness of our Second Opinion Gating mechanism, we performed an ablation experiment comparing our dynamic selection approach to alternative methods, such as averaging logits across layers, random selection, and fixed selection layer 31. The results, presented in Table 4, indicate that our dynamic gating mechanism, which selects the optimal layer based on opinion divergence and consistency, significantly outperforms both fixed and random selection strategies. This experiment confirms that our gating approach can more accurately identify and leverage the appropriate expert layers, thereby reducing hallucinations more effectively.

λ	**1**	**10**	**20**	**50**	**100**	**200**
$\text{CHAIR}_S \downarrow$	20.4	18.0	19.8	20.2	**17.2**	18.4
$\text{CHAIR}_I \downarrow$	7.16	6.22	7.0	7.02	**5.63**	6.12

Table 4: Ablation on the temperature coefficient λ.

Gating Setup	Average	Random	Static	Only_Key	Only_Maj.	MoLE
$\text{CHAIR}_S \downarrow$	22.2	23.2	23.4	20.6	21.2	**17.2**
$\text{CHAIR}_I \downarrow$	7.1	7.8	8.2	7.5	6.9	**5.63**

Table 5: Ablation on the Gating method of SO Expert. "Average" represents the average score of candidate layers. "Random" represents random selection on candidate layers. "Static" represents directly selecting the penultimate layer. "Only_Key" represents selecting the layer with the largest JSD with the final expert on the top-k token logits. "Only_Maj." represents selecting the layer with the smallest JSD with the final expert on the majority token logits.

Influence of the temperature λ. We conducted additional experiments to examine the impact of varying the temperature coefficient λ on the performance of the Prompt Retention Expert. This study aimed to determine the optimal weight that should be assigned to the Prompt Retention Expert at different stages of sequence generation. As shown in Table 5, the results indicate that hallucination metrics reach their optimal values when the temperature coefficient is set to a value that corresponds to $t \approx 100$. This suggests that the influence of the Prompt Retention Expert becomes more critical as the sequence length increases, which is consistent with previous findings on the importance of maintaining attention to the initial prompt. Adjusting the weight dynamically based on sequence progression allows the model to maintain a higher faithfulness to the input prompt, effectively reducing hallucinations in longer sequences. More experiments will be included in the supplementary.

Conclusion

We represent the Mixture of Layer Experts (MoLE) framework to address hallucinations in Large Vision-Language Models (LVLMs). MoLE leverages the collaborative strengths of multiple expert layers within the model to enhance output faithfulness without additional computational costs. Our experiments demonstrated that MoLE outperforms existing methods in reducing hallucinations across various LVLMs.

This work highlights the potential of layer expert collaboration in improving the faithfulness of LVLMs. Future research could explore applying MoLE in diverse multimodal contexts and refining expert selection to further boost performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064, Ant Group, and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. arXiv:2403.00425.

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

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

et al, L. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *arXiv preprint arXiv:2311.16922*.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-Modal Hallucination Control by Visual Information Grounding. arXiv:2403.14003.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. arXiv:2311.17911.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086.

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023b. Contrastive Decoding: Open-ended Text Generation as Optimization. arXiv:2210.15097.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating Object Hallucination in Large Vision-Language Models. arXiv:2305.10355.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024. Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models. arXiv:2311.06607.

Liang, T.; Huang, J.; Kong, M.; Chen, L.; and Zhu, Q. 2024. Querying as Prompt: Parameter-Efficient Learning for Multimodal Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26855–26865.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2024a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. arXiv:2306.14565.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024c. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rawte, V.; Sheth, A.; and Das, A. 2023. A Survey of Hallucination in Large Foundation Models. arXiv:2309.05922.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2019. Object Hallucination in Image Captioning. arXiv:1809.02156.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

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

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *arXiv preprint arXiv:2310.16045*.

Zhao, L.; Deng, Y.; Zhang, W.; and Gu, Q. 2024. Mitigating Object Hallucination in Large Vision-Language Models via Classifier-Free Guidance. arXiv:2402.08680.

Zheng, C.; Li, L.; Dong, Q.; Fan, Y.; Wu, Z.; Xu, J.; and Chang, B. 2023a. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2310.00754*.

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