Mixture of Specialized Vision Experts: Unlocking Complementary Visual Insights for Faithful MLLM Reasoning

Yifei Gao[2†], Liangliang You[1,3†], Jiye Xie[4], Changwei Wang[1,9✉], Kexue Fu[1,9✉], Jingyi Liu[5], Rongtao Xu[6], Zhiqiang Kou[7], Haoran Xu[4], Longxiang Gao[1,9], and Yu Zhang[8]

[1] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology, Jinan, China
[2] Tianjin University, Tianjin, China
[3] North China Electric Power University, Beijing, China
[4] Zhejiang University, Hangzhou, China
[5] Tongji University, Shanghai, China
[6] MBZUAI, Masdar City, Abu Dhabi, UAE
[7] The Hong Kong Polytechnic University, Hong Kong, China
[8] Macquarie University, NSW, Australia
[9] Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China

Abstract. Multimodal Large Language Models (MLLMs) achieve remarkable progress in widespread vision-language tasks, while achieving faithful MLLM reasoning remains a critical challenge: generating content grounded in comprehensive visual evidence without hallucination. This challenge often stems from the inherent perceptual limitations of the single visual encoder design in most MLLMs. While many works have turned to integrating multiple vision experts, existing methods fail to leverage the unique insights of diverse experts due to unspecialized mixture and alignment strategies. To resolve these challenges, we introduce **MoSVE** (**M**ixture of **S**pecialized **V**ision **E**xperts), a holistic framework built upon the core insight of inter-expert complementarity. To achieve fine-grained mixture, we introduce **Query-guided Complementary Clustering**, which selectively preserves text-critical and informative visual perception without redundancy. To cultivate expert specialization, we propose **Complementary Rejection Fine-Tuning**, which explicitly routes hard samples from the anchor MLLM to the most visually-disparate auxiliary expert, enhancing unique insights without homogenization. Extensive experiments demonstrate that MoSVE not only mitigates hallucinations across the POPE, CHAIR and MMVP, but also improves general multimodal reasoning on MMBench. Ultimately, MoSVE provides an advanced and efficient solution for faithful MLLM reasoning.[10]

† Equal contribution. ✉ Emails: changweiwang@sdas.org; fukx@sdas.org

[10] `https://github.com/jerry19h/MoSVE`.

1 Introduction

Multimodal Large Language Models (MLLMs) [1, 3, 6, 18, 28, 29, 34, 42] have recently achieved impressive progress in visual understanding, instruction following and multimodal reasoning. However, despite their success, achieving faithful multimodal reasoning remains a fundamental challenge. MLLMs frequently suffer from hallucinations [4], where the generated content conflicts with the actual visual evidence. This lack of reasoning faithfulness remains a major obstacle to deploying MLLMs in factual and safety-critical applications, necessitating an effective foundation for comprehensive visual grounding.

In MLLMs, the visual subsystem provides critical perceptual evidence for grounding language generation. Therefore, the perception shortcomings are a key cause of unfaithful MLLM reasoning [17, 25, 33]. Most existing MLLMs rely on a single CLIP-based encoder [23]. Despite its strong global semantic alignment, CLIP often suffers from perceptual blind spots in fine-grained visual details such as spatial relations and precise object boundaries [7, 14, 40]. To overcome these inherent limitations, recent studies have turned to integrating multiple vision experts via strategies like token concatenation [2, 30] and auxiliary guidance or refinement [5, 26, 39], enhancing perceptual diversity for faithful reasoning.

Despite the enrichment of visual evidence, current multi-expert frameworks [2, 5, 21, 27, 30, 39] face two critical limitations that hinder faithful MLLM reasoning.(As shown in Fig. 1) *(1) Coarse-grained Expert Mixture*: Most methods conduct simple mixture such as direct concatenation, without exploring multi-expert representation distributions. This coarse-grained combination causes under-utilization of informative visual insights with significant redundancy. *(2) Unspecialized Expert Alignment*: Most approaches align multi-expert spaces only using standard data. This unspecialized alignment often leads to feature conflicts or convergence to homogeneous representations, failing to provide complementary visual insights. This leads to a fundamental question: *How can we effectively specialize and integrate vision experts to unlock their complementary strengths for faithful reasoning?*

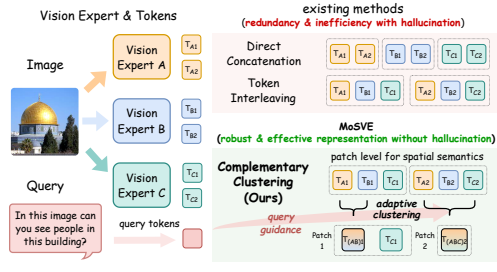

Fig. 1: Comparison of existing multi-expert models and our **MoSVE**. Existing methods introduce coarse-grained integration and unspecialized alignment, while our MoSVE preserves specialized and informative visual semantics.

To address this question, we first conduct a diagnostic representation analysis across diverse vision experts, covering semantic-centric (CLIP [23]), object-oriented (DINOv2 [22]) and structure-aware (SAM-L [13]) viewpoints. Our analysis reveals that these experts exhibit significant **complementarity**: their visual focus shares both overlaps and expert-specific perspectives. Based on this insight, we introduce **MoSVE** (**M**ixture of **S**pecialized **V**ision **E**xperts), a holistic multi-

expert framework for faithful MLLM reasoning. For *fine-grained mixture*, we propose *Query-Guided Complementary Clustering*. Specifically, for each image patch, we perform adaptive hierarchical clustering on multi-expert visual tokens based on query relevance and semantic richness. This mechanism dynamically merges redundant consensus while preserving the unique, informative details required for faithful and comprehensive perception. Furthermore, for *specialized alignment*, we propose *Complementary Rejection Fine-Tuning (CRFT)*. Unlike joint optimization, CRFT identifies failure cases from the anchor CLIP-based MLLM and routes these hard samples to the most visually-disparate auxiliary expert for targeted projector training. This ensures that each expert specializes in overcoming the anchor model's blind spots, thereby maximizing the collective synergy of the multi-expert system.

We conduct extensive experiments of MoSVE across a wide range of faithful MLLM reasoning benchmarks. Our results demonstrate that MoSVE not only mitigates hallucinations on POPE [15], CHAIR [25], and MMVP [30], but also advances general multimodal reasoning capabilities on MMBench [19], outperforming various SOTA MLLMs and multi-expert baselines. Notably, MoSVE achieves this superior performance while maintaining inference efficiency with lightweight token consumption. Our contributions are threefold:

- We introduce **MoSVE** (**M**ixture of **S**pecialized **V**ision **E**xperts), a novel framework based on inter-expert complementarity for faithful MLLM reasoning.
- Our MoSVE introduces two core mechanisms: *Query-Guided Complementary Clustering* for fine-grained mixture of informative visual features, and *Complementary Rejection Fine-Tuning* for specialized alignment via anchor-based sample routing.
- Extensive experiments demonstrate that MoSVE outperforms existing MLLMs and multi-expert baselines on hallucination-centric (POPE, CHAIR, MMVP) and general reasoning (MMBench) benchmarks, providing a novel effective solution for faithful MLLM architectures.

2 Related Work

2.1 Hallucination in MLLMs

Hallucination remains a significant challenge for faithful MLLM reasoning, representing the generated responses present misalignment with the image. Earlier studies on image captioning exposed object hallucinations where models describe nonexistent entities [25]. With the emergence of MLLMs such as MiniGPT-4 [42] and LLaVA [18], hallucinations have expanded beyond object existence to attributes and relations [37]. Recent benchmarks, including POPE [15], CHAIR [25], and MMVP [30], further standardize hallucination evaluation through discriminative and generative protocols. To alleviate hallucination, existing approaches operate at different levels of the MLLM pipeline. Data-centric strategies curate balanced or contrastive instruction datasets [9,31,32,39], vision-centric methods enhance perceptual fidelity via higher-resolution or auxiliary encoders [10,11,37,

43], and alignment-based training or decoding refinement aims to reduce the visual–textual gap [12, 33, 38]. Recent post-processing frameworks, like LURE [41] and Woodpecker [35], further revise outputs through visual verification and factual correction. Despite these advances, hallucination remains a fundamental limitation as MLLMs scale toward more complex and open-ended scenarios.

2.2 Multi-Expert MLLM Framework for Faithful Reasoning

To mitigate hallucination, employing multiple visual encoders to enhance MLLM performance has already been explored. Early frameworks such as BLIP-2 and InstructBLIP [6, 14] pioneered the decoupled design that bridges frozen visual encoders with large language models, laying the foundation for subsequent multi-encoder fusion architectures. LEO [2] adopts hierarchical multi-resolution alignment to maintain consistent reasoning across scales, while MoF [30] introduces a multi-objective fusion strategy balancing semantic and structural representations. LLaVA-HR [21] further integrates high- and low-resolution pathways under a Mixture-of-Resolution Adaptation design, and MARINE [39] employs multi-resolution alignment to explicitly mitigate visual hallucination. Beyond these, multi-stream or auxiliary-perception frameworks such as SPHINX [16] also reflect the growing trend of fusing heterogeneous visual representations. MoVE-KD [5] performs multi-expert knowledge distillation using MoE-routed LoRA [8] and attention-weighted fusion to improve a single encoder. These advances demonstrate the potential of diverse visual perception to overcome the limitations of single-encoder MLLMs.

3 MoSVE: Mixture of Specialized Vision Experts

In this section, we conduct a diagnostic analysis of multi-expert representations to confirm our core insight: **inter-expert complementarity** (Sec 3.1). Building on this insight, we present MoSVE (Mixture of Specialized Vision Experts). MoSVE achieves robust and comprehensive multimodal understanding via two core mechanisms: Query-Guided Complementary Clustering as adaptive mixture (Sec 3.2) to integrate heterogeneous expert features without redundancy; Complementary Rejection Fine-Tuning (CRFT) as specialization enhancement (Sec 3.3) to cultivate unique expert insights. Together, MoSVE overcomes single-encoder limitations via faithful perception beneficial for multimodal reasoning.

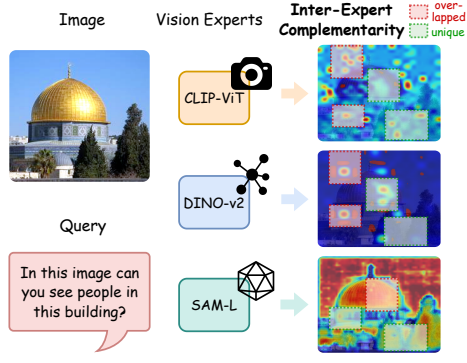

Fig. 2: Visualization of inter-expert complementarity via attention heatmaps. For the same input image, different vision experts exhibit complementary focus areas.

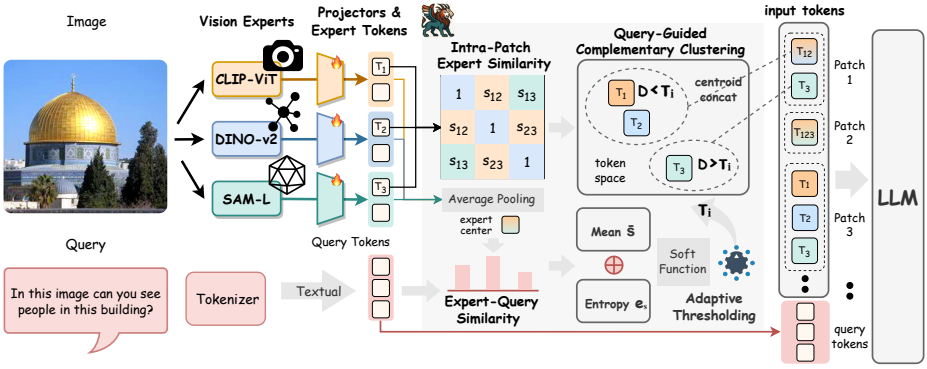

Fig. 3: The detailed architecture of our Query-Guided Complementary Clustering. For each image patch, our framework computes a adaptive threshold based on the semantic relevance and richness between the multi-expert visual tokens and input query. This adaptive threshold guides a patch-level hierarchical clustering process that merges redundant expert features while preserving complementary tokens, resulting in context-aware and robust visual representation for faithful MLLM reasoning.

3.1 Observation of Inter-Expert Complementarity

To comprehensively understand the capabilities and focusing regions of different vision experts, we use multiple widely-used visual encoders (CLIP [23], DINOv2 [22], and SAM-L [13]) to extract features from a single image and conduct a diagnostic analysis of their feature distributions. To this end, we generate attention heatmaps by aggregating self-attention scores from the last layer in Vision Transformers (ViT) and projecting them into the original spatial space. This attention heatmap reflects the most important visual details within their respective representations.

As shown in Fig. 2, our selected vision experts exhibit distinct perception patterns. CLIP mainly pays attention to high-level semantic regions related to the input query, showing strong global cross-modal perception. DINOv2 can highlight object attributes and local features without textual guidance, revealing strong object-oriented perception. SAM-L focuses on instance-level shapes and precise boundaries due to its segmentation-driven training. These attention patterns indicate that visual experts exhibit significant **complementarity**: they share focus on common salient regions while also providing their own independent perspectives. Therefore, this motivates a two-fold requirement: adaptively mixing consistent visual evidence for informative perception, while preserving and actively cultivating specialized expert insights. This complementary paradigm constructs the foundation of MoSVE, enabling comprehensive and reliable visual representations for faithful MLLM reasoning.

3.2 Query-Guided Complementary Clustering

Building upon the core insight of complementarity, we require an effective mixture mechanism that can intelligently integrate diverse vision experts. Existing token mixture approaches like direct concatenation or simple averaging fail to distinguish between redundant consensus and valuable, unique perspectives. We recognize that merging similar information while preserving distinct outliers aligns perfectly with the typical goal of clustering algorithms [24]. This motivates our patch-level clustering approach for informative visual representation from multiple experts, as illustrated in Fig. 3. Meanwhile, the patch-level mixture can also maintain the spatial relationships in visual space, facilitating faithful and accurate perception.

For the i-th image patch, we collect the expert tokens given by corresponding encoders and projectors as $\mathcal{V}_i = \{\mathbf{v}_i^{(k)} \in \mathbb{R}^d \mid k = 1, \ldots, N_e\}$, where d is the token embedding dimension and N_e is the number of vision experts.

Inter-Expert Similarity To evaluate consistency among these multi-expert tokens, we compute a pairwise cosine similarity matrix S_i for the i-th patch:

$$S_i(m,n) = \frac{\langle \mathbf{v}_i^{(m)}, \mathbf{v}_i^{(n)} \rangle}{\|\mathbf{v}_i^{(m)}\|_2 \, \|\mathbf{v}_i^{(n)}\|_2}, \tag{1}$$

and define the corresponding distance as $D_i(m,n) = 1 - S_i(m,n)$. Those tokens with small distance indicate similar or redundant information, while larger distance implies complementary visual evidence.

Query-Guided Adaptive Thresholding For MLLM reasoning, the faithful responses without hallucination rely on comprehensive visual understanding based on the input query. Therefore, we propose a query-guided adaptive thresholding mechanism to dynamically control the clustering strength. We derive this adaptive threshold from two following key metrics:

- *Cross-Modal Relevance*: Visual regions relevant to the textual query are essential for accurate grounding and faithful answering. Therefore, if the expert tokens exhibit cross-modal alignment, the mixture should be conservative to preserve valuable evidence.
- *Expert-Query Semantic Divergence*: When image patches have rich connections with multiple aspects of the input query, it is necessary to retain different expert perspectives for visual enhancement.

Specifically, to measure these factors for the i-th patch, we first aggregate its expert features via mean pooling to get an expert center $\bar{v}_i = \frac{1}{N_e} \sum_{k=1}^{N_e} v_i^{(k)}$. We then compute its cosine similarity with each of the N_t text tokens from the input query, yielding a similarity vector $s_i \in \mathbb{R}^{N_t}$. The cross-modal relevance $\bar{s}_i$

and inter-expert semantic divergence e_i are then calculated as:

$$\begin{aligned} \bar{s}_i &= \frac{1}{N_t} \sum_{t=1}^{N_t} s_i^{(t)} \\ e_i &= -\sum_{t=1}^{N_t} p(s_{i,t}) \log p(s_{i,t}), \ \text{ where } p = \text{Softmax}(s_i) \end{aligned} \tag{2}$$

Based on these two metrics, we design a soft function to control the clustering threshold adaptively. The final threshold τ_i for the i-th patch is formulated as:

$$\tau_i = 1 - \sigma\big(\alpha\, \bar{s}_i + \beta\, e_i\big) \tag{3}$$

where σ denotes the sigmoid function, α and β are hyperparameters which balance the contribution of relevance and divergence. Through our carefully-designed adaptive threshold mechanism, we ensure that achieving expert complementarity is tightly integrated with the input query, facilitating targeted and comprehensive visual understanding for specific scenarios.

Token Mixture via Cluster Centroid For the i-th image patch, given the distance matrix D_i and adaptive threshold τ_i, we perform hierarchical clustering over the expert tokens $\mathcal{V}_i$. Visual tokens whose pairwise distances fall below τ_i are assigned to the same cluster $\mathcal{C}_{i,m}$. To mix the redundant evidence within each cluster, we compute the centroid of all member tokens. Formally,

$$\begin{aligned} \hat{\mathbf{v}}_{i,m} &= \frac{1}{|\mathcal{C}_{i,m}|} \sum_{\mathbf{v} \in \mathcal{C}_{i,m}} \mathbf{v}, \\ \hat{\mathbf{V}} &= [\, \hat{\mathbf{v}}_{1,1}, \ldots, \hat{\mathbf{v}}_{i,m}, \ldots], \end{aligned} \tag{4}$$

where $\hat{\mathbf{V}}$ is the fused visual token sequence formed by concatenating all cluster centroids across spatial positions.

In conclusion, our cluster-based mixture mechanism preserves complementary evidence across diverse experts and improves query-guided visual semantics, promoting faithful MLLM reasoning with token and parameter efficiency.

3.3 Complementary Rejection Fine-Tuning

Although mainstream visual encoders capture diverse semantics, simply projecting them into the LLM space via standard joint fine-tuning may face feature conflicts or cause representation homogenization, leading to suboptimal multimodal perception. Therefore, we freeze the vision experts and the LLM backbone, focusing on training a dedicated projector for each expert.

To fully unlock the potential of our mixture mechanism and prevent homogeneous convergence, we propose *Complementary Rejection Fine-Tuning (CRFT)*. Instead of standard joint training on all data, CRFT explicitly specializes each auxiliary expert by producing unique visual tokens to supplement helpful insights. This targeted specialization preserves complementary evidence and effectively corrects the inherent biases of the existing vision encoder.

Anchor-Based Rejection Sampling Since CLIP serves as the primary vision encoder in most MLLMs (e.g., LLaVA), we utilize a CLIP-based MLLM as the anchor to identify cases of unfaithful reasoning caused by single-encoder visual biases. Specifically, across the training data, we use the anchor model to generate responses and verify their correctness. Samples yielding incorrect or hallucinated responses are collected into a negative pool ($\mathcal{P}_{neg}$). This pool represents hard cases where CLIP-based visual representations are insufficient, necessitating complementary visual perception to achieve faithful reasoning.

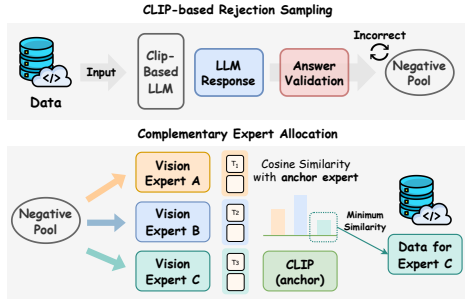

Fig. 4: Workflow of Complementary Rejection Fine-Tuning (CRFT). Anchor-based rejection sampling constructs a negative pool from anchor failures. Each sample is then routed to the auxiliary expert with maximal feature discrepancy, encouraging complementary specialization.

Complementary Expert Routing
To fully cultivate expert specialization, we hypothesize that the auxiliary visual expert exhibiting the greatest feature discrepancy from the anchor presents the highest potential to correct the anchor model's perceptual biases. As shown in Fig. 4, for a given hard image in $\mathcal{P}_{neg}$, we first obtain a global image-level representation $\bar{\mathbf{v}}^{(k)}$ for each expert k via average pooling across its visual tokens. We then compute the cosine similarity between the anchor expert's representation ($\bar{\mathbf{v}}^{(anchor)}$) and each auxiliary expert's representation. Crucially, the hard sample is exclusively routed to the specific training subset $\mathcal{D}_k$ of the auxiliary expert k^* that minimizes this similarity:

$$k^* = \arg\min_{k \in Aux} S(\bar{\mathbf{v}}^{(anchor)}, \bar{\mathbf{v}}^{(k)}), \tag{5}$$

where S denotes cosine similarity. This explicit discrepancy-driven routing is the fundamental advantage of CRFT over standard fine-tuning: it effectively incentivizes the auxiliary experts to specialize in resolving specific multimodal reasoning biases.

Two-Stage Training for Expert Specialization and Mixture We implement CRFT through a two-stage training paradigm to first cultivate specialization and then align them for multimodal reasoning. In the first stage, we independently train the lightweight projector of each auxiliary expert using its exclusively routed hard-sample subset ($\mathcal{D}_k$). This explicit routing guides each projector to extract highly specialized, complementary visual evidence. In the second stage, we combine the general training data and perform supervised fine-tuning on all projectors and the LLM backbone concurrently. This two-stage

process ensures that vision experts develop unique perceptual strengths and collaborate complementarily, yielding robust and informative visual representations crucial for faithful MLLM reasoning.

4 Experiments

4.1 Experimental Setup

Training Dataset and Negative Pool Construction Our training data is collected from a comprehensive multimodal conversation dataset EAGLE-1.8M[11]. To prevent data leakage, we remove any samples overlapping with the LLaVA-1.5 SFT dataset. This difference set is used for negative pool construction ($\mathcal{P}_{neg}$) and subsequent fine-tuning.

For correctness verification during CRFT, we adopt a strict strategy based on answer verifiability: for explicit tasks (e.g., yes/no, choice), we employ exact answer matching; for free-form generations, we extract ground-truth object attributes (e.g., existence, color, shape) by spaCy toolbox[12] and check whether the response ignores specified objects or describes conflicting attributes as our evaluation criterion.

Benchmarks POPE [15]. To evaluate object hallucination, we utilize the widely-used POPE benchmark. POPE assesses a model's ability to confirm the presence or absence of specific objects through binary "*Yes/No*" questions. We report overall accuracy and F1 score as comprehensive POPE evaluation.
CHAIR [25]. The CHAIR benchmark is a standard for quantifying object hallucination in generated image descriptions. For each image, CHAIR constructs a set of ground-truth objects, and any object mentioned in the generated description but not in this ground-truth set is regarded as a hallucination object. CHAIR provides metrics based on two dimensions: sentence and instance level, denoted as C_S and C_I, respectively.

$$\begin{aligned} \mathrm{CHAIR}_S &= \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|} \\ \mathrm{CHAIR}_I &= \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all annotated objects}\}|} \end{aligned} \tag{6}$$

MMVP [30]. MMVP is uniquely designed to expose systematic visual failures by using "*CLIP-blind pairs*" (i.e., image pairs which are highly similar in CLIP space but have clear visual differences). It requires MLLMs to distinguish pairwise images with similar semantics. Overall, MMVP directly requires the MLLM to perceive visual patterns (e.g., orientation, quantity, spatial relations) where a single encoder may be insufficient, validating the effectiveness of our complementary design in MoSVE.

[11] https://huggingface.co/datasets/shi-labs/Eagle-1.8M
[12] https://github.com/explosion/spaCy

Table 1: Comparison of our MoSVE and existing MLLM / multi-expert baselines on POPE, CHAIR, MMVP and MMBench. *Expert Paradigm* represents how MLLMs uses their vision experts, including training and inference strategies. *Image Tokens* denotes the number of visual tokens required to answer questions. The **bold** and underlined styles respectively indicate the top and second-best results.

Model	Expert Paradigm	LLM Backbone	Image Tokens	POPE		CHAIR		MMVP	MMB
				Accu↑	F1↑	C_S↓	C_I↓	P-Accu↑	Accu↑
LLaVA-1.5	Single-stage	Vicuna-7B	576	83.1	82.9	45.6	16.3	22	64.3
LLaVA-1.5	Single-stage	Vicuna-13B	576	83.9	83.6	41.3	14.2	22	67.7
mPLUG-Owl	Two-stage	LLaMA-7B	64	83.2	82.9	39.9	14.8	–	64.5
SPHINX	Multi-stage	Vicuna-13B	2880	81.1	80.1	42.7	15.6	23.5	65.9
LLaVA-NeXT	Multi-stage	Vicuna-v1.5-7B	2880	85.5	85.2	36.9	12.4	39	67.4
LLaVA-NeXT	Multi-stage	Hermes-Yi-34B	2880	86.1	86.0	35.8	12.0	44	-
MARINE	Guidance	Vicuna-v1.5-7B	–	84.3	83.8	37.5	12.7	42	65.7
MoF	Expert-fusion	Vicuna-v1.5-7B	512	84.3	84.2	39.8	15.2	28	66.2
LLaVA-HR	Multi-scale	Vicuna-v1.5-7B	1024	85.9	85.8	38.0	13.2	52	68.6
MoVE-KD	Distillation	Vicuna-v1.5-7B	577	86.3	-	-	-	37	67.6
LEO	Multi-stage	InternLM2-7B	3584	86.5	86.3	31.2	12.1	46	67.8
MoSVE	**Complementary**	Vicuna-v1.5-7B	**576**	**86.9**	**86.5**	**27.1**	12.2	**57**	68.3

MMBench [19]. MMBench is a comprehensive multiple-choice benchmark for evaluating general multimodal reasoning across diverse visual tasks, including object recognition, attribute understanding, spatial reasoning, and OCR. We use MMBench to verify that our MoSVE achieves not only performs faithful reasoning, but also showcases superior general multimodal capabilities.

Implementation Details We implement MoSVE on top of the standard LLaVA architecture. To instantiate our mixture of specialized experts, we integrate the CLIP-L/14-336 (as the anchor), DINOv2-L/14 and SAM-L/16. For our query-guided clustering, the adaptive thresholding parameters were empirically set to $\alpha = 0.45$ and $\beta = 0.75$.

The two-stage training settings are as follows. First, we train expert-specific projectors using a learning rate of 1e-4. Subsequently, the full model undergoes end-to-end instruction tuning for a single epoch with a lower learning rate of 2e-5. For both stages, we use batch size of 64 and employ the AdamW [20] optimizer with a cosine decay schedule. We conduct training and evaluation on 8 and 1 NVIDIA A100 GPU(s), respectively.

4.2 Main Results

We conduct comprehensive comparison of our MoSVE against a wide range of existing MLLMs and multi-expert baselines on POPE, CHAIR, MMVP and MMBench. Tab.1 presents the main results, where our MoSVE is evaluated on its ability to perform robust visual reasoning and maintain computational efficiency.

Comparison of Base MLLMs The first part of our comparison includes prominent MLLM baselines such as LLaVA-1.5, mPLUG-Owl, SPHINX and LLaVA-NeXT. These models, while powerful, demonstrate the limitations of relying on single-expert paradigms. For instance, even the strongest baseline, LLaVA-NeXT with the large Hermes-Yi-34B [36] backbone, achieves a CHAIR_S score of 35.8 and MMVP accuracy of only 44. This indicates that merely scaling the LLM backbone or blindly increasing visual tokens (e.g., SPHINX uses 2880 tokens) is insufficient to resolve the root cause of unfaithful reasoning. Without specialized visual experts, standard MLLMs struggle to ground their generation in accurate visual evidence.

In contrast, MoSVE significantly outperforms all standard MLLM baselines across every benchmark. Compared to the strongest LLaVA-NeXT (Hermes-Yi-34B), our method achieves 8.7% performance gain on CHAIR_S (27.1 against 35.8) and an impressive 13% improvement on the challenging MMVP benchmark (57 against 44). Notably, our superior performance is achieved using a lightweight Vicuna-v1.5-7B backbone. This highlights that complementary expert specialization and mixture are more effective solutions for faithful MLLM reasoning than simple model scaling.

Comparison of Multi-Expert Frameworks [13] We emphasize the comparison between our MoSVE and existing SOTA multi-expert frameworks, including MARINE, MoF, LLaVA-HR, MoVE-KD and LEO. These methods represent advanced attempts to leverage diverse visual information for multimodal reasoning. While they generally outperform the base MLLMs, existing multi-expert strategies suffer from the limitations identified in our analysis: *inefficient mixture and unspecialized alignment*.

For instance, MoF achieves limited accuracy (28 on MMVP), suggesting that its coarse-grained token interleaving fails to effectively coordinate diverse perceptual patterns. Similarly, MoVE-KD (37 on MMVP) relies on distillation without explicit specialization, leading to less robust visual representations. LEO requires a massive budget of 3584 image tokens caused by direct token concatenation.

As shown in Tab.1, we achieve the best performance on POPE accuracy (86.9) and F1 (86.5), CHAIR_S (27.1), and MMVP score (57), demonstrating superior ability to handle fine-grained patterns and overcome existing visual shortcomings. Critically, this leading performance is achieved with remarkable efficiency. Our model uses only 576 visual tokens, which is 6.2 times fewer than LEO (3584) and nearly half that of LLaVA-HR (1024). These robust results prove that our query-guided complementary clustering serves as an fine-grained and efficient mixture mechanism, and our CRFT strategy incentivizes expert-specific visual insights, achieving faithful multimodal understanding and reasoning.

Evaluation of General Multimodal Reasoning Furthermore, MoSVE achieves excellent performance on the general multimodal reasoning benchmark MM-Bench (68.3). The results demonstrate that by fixing perceptual blind spots,

[13] We evaluate the per-sample inference latency in Tab.5.

MoSVE establishes a more comprehensive visual foundation that benefits general multimodal reasoning tasks across practical scenarios.

4.3 Ablation Analysis of MoSVE components

Mixture Stage Analysis: Query-Guided Complementary Clustering To validate the effectiveness of our proposed query-guided clustering mechanism, we compare our method against two critical variants:

- *Fixed + CRF*: This version replaces our adaptive threshold with an empirically optimized fixed threshold ($\tau = 0.45$). It serves as a strong baseline to determine if a static non-contextual clustering approach is sufficient.
- *Upper Limit*: This variant bypasses clustering module and feeds all specialized expert tokens (1536 tokens in total) to LLM backbone. This setting has relative high cost, but it serves as theoretical ceiling by preserving the maximum visual information while still benefiting from our CRFT strategy.

The results in Tab.2 clearly demonstrate the superiority of query-guided clustering. First, MoSVE (QG+CRF) significantly outperforms the Fixed variant (consistent gains on POPE, CHAIR_S and MMVP), confirming that dynamically adapting mixture strength based on query semantics is crucial for preserving text-critical evidence.

Table 2: Ablation results of our query-guided (QG) complementary clustering. For fair comparison, we conduct our CRFT for all variants. The upper limit is indicated in *italics*.

Method	Image Tokens	POPE↑	$\text{CHAIR}_S \downarrow$	MMVP↑
Fixed+CRF	600	85.1	34.0	48
QG+CRF	**576**	**86.9**	**27.1**	**57**
All visual tokens+CRF	1536	*87.2*	*25.1*	*59*

Second, compared to the theoretical upper limit with full-scale tokens, MoSVE retains 96.6% of the performance (57 against 59 on MMVP) while using 2.67 times fewer image tokens (576 against 1536). This demonstrates that our query-guided clustering achieves a remarkable balance, compressing redundancy while retaining the complementary insights essential for faithful MLLM reasoning.

Specialization Stage Analysis: Expert Diversity and CRFT We investigate the contributions of diverse vision experts and our novel CRFT in Tab.3.

First, simply adding experts (from C to C+D+S) improves results progressively (from 22 to 42 on MMVP), indicating that diverse experts provide complementary visual information essential for robust visual grounding.

Crucially, when we apply our CRFT strategy, our

Table 3: Ablation results of diverse vision experts and our CRFT. C, D and S denote CLIP, DINOv2 and SAM-L experts, respectively.

Methods	CRFT	POPE↑	$\text{CHAIR}_S \downarrow$	MMVP↑
C	✗	83.1	45.6	22
C+D	✗	84.5	38.8	38
C+S	✗	84.2	39.5	36
C+D+S	✗	85.3	36.8	42
MoSVE	✓	**86.9**	**27.1**	**57**

MoSVE achieves a substantial 15% on MMVP (from 42 to 57) and 9.7% on CHAIR_S (from 36.8 to 27.1). This significant improvement highlights the specialization cultivated by CRFT is beneficial to unlock full potential for multi-expert system, thus achieving superior multimodal reasoning performance.

Routing Mechanism Analysis: Complementary Expert Routing To verify the effectiveness of the expert routing mechanism in our CRFT, we design two variants as follows:

- *Random Routing*: assigns the negative samples to the vision experts equally at random.
- *Standard SFT*: Fine-tunes all expert projectors on the entire negative pool via joint training.

As shown in Tab.4, the weak result of random routing proves that blind assignment confuses the vision experts. Standard SFT underperforms MoSVE consistently on POPE, CHAIR and MMVP, indicating that indiscriminately training hard samples is insufficient to develop unique expert strengths. In contrast, MoSVE's complementary routing explicitly assigns failures to the most potential expert, maximizing specialized gains. This indicates that data construction and utilization strategies are the key elements which enable full capabilities of the multi-expert system.

Table 4: Ablation results of complementary routing mechanism in CRFT.

Method	**POPE**↑	**CHAIR**$_S$ ↓	**MMVP**↑
Random	85.7	30.8	50
Standard SFT (no routing)	86.5	29.2	54
CRFT	**86.9**	**27.1**	**57**

4.4 Inference Latency Analysis

Tab.5 reports the inference throughput and latency on POPE. The results demonstrate that our MoSVE achieves faithful multimodal reasoning without significantly increasing inference resource consumption. This is attributed to the complementary clustering that compresses visual tokens into informative and robust representations, which are both efficient and beneficial for multimodal reasoning.

Table 5: Inference latency comparison on POPE.

Method	**Throughput** (samples/s)	**Latency** (s/sample)	**POPE Accu**↑
MARINE	0.67	1.49	84.3
MoF	0.70	1.44	84.3
LLaVA-HR	0.62	1.63	85.9
MoVE-KD	0.77	1.31	86.3
LEO	0.29	3.49	86.5
MoSVE	0.68	1.49	**86.9**

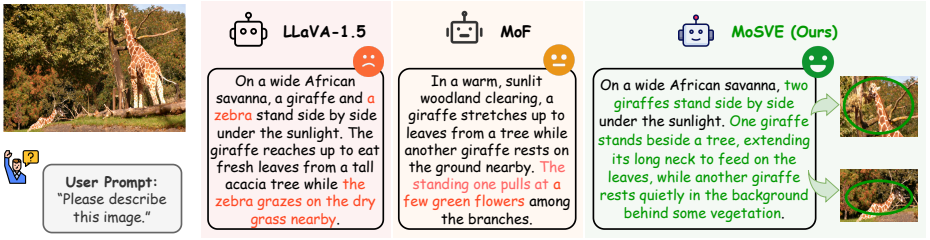

Fig. 5: A qualitative case study comparing MoSVE with the LLaVA-1.5 and MoF baselines. While LLaVA-1.5 and MoF respectively hallucinate non-existent zebra and flowers (marked in red), MoSVE correctly identifies both giraffes (marked in green) with a factually grounded detailed description, demonstrating that our specialization and mixture paradigm effectively resolves perceptual blind spots for faithful reasoning.

4.5 Qualitative Analysis

To provide an intuitive understanding of MoSVE's effectiveness, we present a qualitative comparison in Fig. 5.

The single-encoder LLaVA-1.5 fails to perceive the second, partially occluded giraffe. Due to perceptual blind spots, it hallucinates a "zebra" (commonly associated with giraffes) and imagines "grazing on dry grass". Notably, the multi-expert baseline MoF, despite having access to additional visual encoders, also performs unfaithful multimodal reasoning. MoF hallucinates "green flowers", suggesting that its coarse-grained token interleaving introduces visual noise instead of clarifying the scene.

In contrast, MoSVE accurately identifies both giraffes and correctly describes their distinct postures (one standing, one feeding). This success highlights the synergy of our design: for *expert mixture*, our query-guided clustering efficiently integrates informative insights without introducing noise; for *expert specialization*, our CRFT strategy successfully specializes the vision experts to capture comprehensive details like object boundaries. This case highlights MoSVE's ability to create robust and informative visual representations, beneficial for faithful multimodal reasoning.

5 Conclusion

In this work, we address the critical challenge of faithful MLLM reasoning by overcoming the perceptual blind spots inherent in single-encoder architectures. We introduce MoSVE (Mixture of Specialized Vision Experts), a holistic framework driven by the core insight of inter-expert complementarity. MoSVE realizes this through two collaborative mechanisms: Query-Guided Complementary Clustering serves as an efficient mixture strategy for informative representations, while Complementary Rejection Fine-Tuning (CRFT) acts as a specialization enhancement to route hard samples for targeted expert alignment. Extensive

experiments demonstrate that MoSVE achieves superior performance on hallucination benchmarks (POPE, CHAIR, MMVP) and general multimodal capabilities (MMBench), all while maintaining high inference efficiency. Ultimately, MoSVE presents a novel multi-expert paradigm, proving that the path toward faithful and robust MLLMs lies not only in simply scaling visual perception patterns, but also in the intelligent mixture of specialized and complementary perceptual insights.

6 Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2025YFE0216800, Key R&D Program of Shandong Province, China (No. 2025KJHZ013), Shandong Provincial Natural Science Foundation for Young Scholars Program (No. ZR2025QC1627,ZR2024QF110), Shandong Provincial University Youth Innovation and Technology Support Program No.2022KJ291, Qilu University of Technology (Shandong Academy of Sciences) Major Project under Grant (No. 2025ZDZX02), Taishan Scholars Program (No.TSQN202507241, TSQN202408245), and National Natural Science Foundation of Grant 62501340.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
2. Azadani, M.N., Riddell, J., Sedwards, S., Czarnecki, K.: Leo: Boosting mixture of vision encoders for multimodal large language models. arXiv preprint arXiv:2501.06986 (2025)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), `https://arxiv.org/abs/2308.12966`
4. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
5. Cao, J., Zhang, Y., Huang, T., Lu, M., Zhang, Q., An, R., Ma, N., Zhang, S.: Movekd: Knowledge distillation for vlms with mixture of visual encoders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19846–19856 (2025)
6. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
7. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14375–14385 (2024)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR (2022)
9. Hu, H., Zhang, J., Zhao, M., Sun, Z.: Ciem: Contrastive instruction evaluation method for better instruction tuning. arXiv preprint arXiv:2309.02301 (2023)
10. Huang, R., Ding, X., Wang, C., Han, J., Liu, Y., Zhao, H., Xu, H., Hou, L., Zhang, W., Liang, X.: Hires-llava: Restoring fragmentation input in high-resolution large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29814–29824 (2025)
11. Jain, J., Yang, J., Shi, H.: Vcoder: Versatile vision encoders for multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27992–28002 (2024)
12. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27036–27046 (2024)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
14. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)

15. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.20, https://aclanthology.org/2023.emnlp-main.20/
16. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
17. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
19. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
21. Luo, G., Zhou, Y., Zhang, Y., Zheng, X., Sun, X., Ji, R.: Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=1EnpStvBU8
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024), https://openreview.net/forum?id=a68SUt6zFt
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
24. Rohlf, F.J.: Adaptive hierarchical clustering schemes. Systematic Biology **19**(1), 58–82 (1970)
25. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4035–4045. Association for Computational Linguistics, Brussels, Belgium (2018)
26. Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., Chadha, A.: A comprehensive survey of hallucination in large language, image, video and audio foundation models. In: Al-Onaizan, Y., Bansal, M., Chen, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024. pp. 11709–11724. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.FINDINGS-EMNLP.685, https://doi.org/10.18653/v1/2024.findings-emnlp.685
27. She, Y., Wu, H.: Fusion to enhance: Fusion visual encoder to enhance multimodal language model. arXiv preprint arXiv:2509.00664 (2025)

28. Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Zhao, Y., Huang, D., Yin, H., Sapra, K., Yacoob, Y., Shi, H., Catanzaro, B., Tao, A., Kautz, J., Yu, Z., Liu, G.: Eagle: Exploring the design space for multimodal llms with mixture of encoders. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=Y2RW9EVwhT
29. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
30. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9568–9578 (2024)
31. Tu, Y., Hu, R., Sang, J.: Ode: Open-set evaluation of hallucinations in multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19836–19845 (2025)
32. Wu, J., Shi, Z., Wang, S., Huang, J., Yin, D., Yan, L., Cao, M., Zhang, M.: Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025)
33. Yang, Z., Luo, X., Han, D., Xu, Y., Li, D.: Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10610–10620 (2025)
34. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
35. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
36. Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Wang, G., Li, H., Zhu, J., Chen, J., et al.: Yi: Open foundation models by 01. ai. arXiv preprint arXiv:2403.04652 (2024)
37. Zhai, B., Yang, S., Xu, C., Shen, S., Keutzer, K., Li, M.: Halle-switch: Controlling object hallucination in large vision language models. arXiv e-prints, pp. arXiv–2310 (2023)
38. Zhang, C., Wan, Z., Kan, Z., Ma, M.Q., Stepputtis, S., Ramanan, D., Salakhutdinov, R., Morency, L., Sycara, K.P., Xie, Y.: Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=tTBXePRKSx
39. Zhao, L., Deng, Y., Zhang, W., Gu, Q.: Mitigating object hallucination in large vision-language models via image-grounded guidance. In: Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025)
40. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16793–16803 (2022)
41. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. In: The Twelfth International Conference on Learning Representations, ICLR 2024,

Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=oZDJKTlOUe

42. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=1tZbq88f27
43. Zhu, S., Dong, W., Song, J., Wang, Y., Guo, Y., Zheng, B.: FILA: Fine-grained vision language models. arXiv preprint arXiv:2412.08378 (2024)