Looking Back and Forth: Cross-Image Attention Calibration and Attentive Preference Learning for Multi-Image Hallucination Mitigation

Xiaochen Yang[1,2†], Hao Fang[1†], Jiawei Kong[1], Yaoxin Mao[3], Bin Chen[4#], and Shu-Tao Xia[1]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Harbin Institute of Technology, Weihai, China
[3] Beijing Institute of Technology, Beijing,China
[4] Harbin Institute of Technology, Shenzhen, China
2022211854@stu.hit.edu.cn, fangh25@mails.tsinghua.edu.cn

Abstract. Although large vision-language models (LVLMs) have demonstrated remarkable capabilities, they are prone to hallucinations in multi-image tasks. We attribute this issue to limitations in existing attention mechanisms and insufficient cross-image modeling. Inspired by this, we propose a structured hallucination mitigation framework involving **C**ross-Image **A**ttention calibration and **P**reference **L**earning **(CAPL)**. CAPL explicitly enhances inter-image interactions at the architectural level while reinforcing reliance on genuine cross-image evidence during training, thereby improving the model's perception and modeling of cross-image associations. Specifically, we (i) introduce a selectable image token interaction attention mechanism to establish fine-grained cross-image entity alignment and information flow; (ii) design a cross-image modeling–based preference optimization strategy that contrasts reasoning outcomes under full inter-image interaction and those obtained when images are mutually invisible, encouraging the model to ground its predictions in authentic visual evidence and mitigating erroneous inferences driven by textual priors. Experimental results demonstrate that CAPL consistently improves performance across multiple model architectures, achieving stable gains on both multi-image hallucination and general benchmarks. Notably, performance on single-image visual tasks remains stable or slightly improves, indicating strong generalization capability. The code is available at https://github.com/yyyxcleo/CAPL.

Keywords: Vision-Language Models · Multi-Image Hallucination · Attention Correction Mechanism

[†]Equal contribution.
[#]Corresponding author.

1 Introduction

Recently, large-scale visual language models (LVLMs) have achieved significant progress in single-image visual question answering, description generation, and reasoning tasks. With the growing demand for multi-image inputs in real-world applications, such as multi-view comparison and cross-image information integration, models are required not only to understand individual images but also to establish relationships and consistency across images. Accordingly, models like Idefics [22] and Qwen-VL [45] have begun to support multi-image inputs and demonstrate capabilities in multi-image understanding tasks. However, LVLMs [33] exhibit serious hallucination issues in downstream tasks [31], *i.e.*, generate plausible yet completely factually incorrect answers. Most existing mitigation methods focus on single-image scenarios and reduce hallucinations through decoding stratgies [10,17] or alignment training [18]. While several decoding-based studies [26,42] have also been proposed to handle the multi-image hallucination, their improvements remain limited since they only adjust the decoding distribution via local structural modifications, without fundamentally enhancing the model's cross-image interactions. Another line of research [37, 48] incorporates delicate training to enhance multi-image reasoning capability for hallucination reduction. However, they still treat each image as an independent context without explicitly modeling inter-image relationships, making it difficult to effectively address hallucinations in multi-image tasks.

Multi-image tasks require the model not only to understand the individual semantics of each image but also to establish symmetric and stable relational modeling across images. Existing Transformer-based autoregressive LVLMs process text and image tokens under a unified causal attention framework [46]. In this setup, multi-image inputs are processed sequentially, *i.e.*, later images can attend to earlier ones, while earlier images have no access to representations of later images, thereby introducing an inherent position bias. This asymmetric information flow undermines semantic equivalence in cross-image relational modeling, making it difficult for the model to establish stable associations across images. Under such an order-biased interaction pattern, the cross-image information flow exhibits a predominantly unidirectional propagation, which hinders the model from establishing symmetric and stable relational alignment at the image-token level. As a result, multi-image reasoning may degenerate into superficial correlation matching over text tokens, rather than genuine relational modeling grounded in visual ones. As illustrated by the first response in Fig. 1(c), model predictions in such cases become increasingly susceptible to language priors [13] and autoregressive generation biases due to insufficient cross-image visual interactions, ultimately leading to unreliable relational inferences.

To address the aforementioned issues, we propose a novel mitigation framework comprising of **C**ross-Image **A**ttention and **P**reference **L**earning **(CAPL)**. Specifically, we first introduce a selective cross-image token mutual attention mechanism to reduce positional bias, as shown in Fig. 1(a). This strategy enables key tokens from different images to interact with each other explicitly during inference. By selectively reactivating the attention between critical cross-

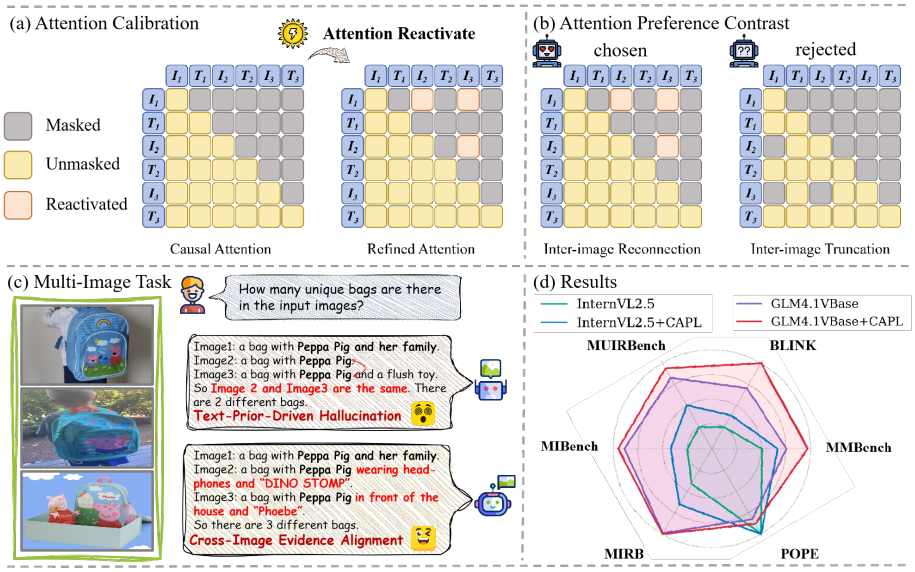

Fig. 1: Overview of our training framework and evaluation. (a) Attention Calibration: Reactivates cross-image token attention on top of causal attention. (b) Attentive Preference Contrast: Generates positive and negative attention for preference learning via inter-image reconnection and truncation. (c) Multi-Image Task: Illustrates a consistency question; incorrect answers ignore inter-image distinctions while correct ones leverage key inter-image information. (d) Evaluation: Our framework perform well on multi-image hallucination and general tasks, while preserving single-image capabilities.

image token pairs, CAPL facilitates structured inter-image information exchange and promotes a more balanced bidirectional attention flow, which effectively alleviates relational modeling limitations induced by the unidirectional causal attention mechanism. Next, we note that LVLMs are generally pretrained under the causal attention paradigm and may lack adaptability to the proposed bidirectional cross-image attention. As a result, directly adapting them to this design may not fully unleash the potential benefits. To handle this challenge, we propose to construct dedicated samples to conduct further training with the proposed attention technique, which enables better adaptation to this mutual interaction paradigm. An intuitive training approach is to perform supervised fine-tuning (SFT) [4] on the LVLM using high-quality data samples. However, we reflect that the SFT objective utilizes only positive samples, without considering and penalizing the models' inherent hallucinated responses, which may fail to effectively suppress their own erroneous reasoning patterns.

Based on these analyses, we introduce an attentive preference alignment approach based on Direct Preference Optimization (DPO) [38]. To obtain preferred non-hallucinated samples, we equip the LVLMs with the proposed attention correction mechanism and query them with different questions. The generated

responses are further improved by incorporating feedback from an advanced Qwen3 model. For rejected sample construction, we aim to encourage the LVLM to fully expose its hallucination behaviors. Concretely, we draw inspiration from the causal attention mechanism in multi-image scenarios from an inverse perspective. In the original causal attention, information flows unidirectionally from earlier to later images. To generate more hallucinated negative responses, we exploit this limitation by completely truncating all cross-image attention connections, enforcing representational independence across all images. As a result, the model is forced to rely solely on individual images and language priors during inference, increasing the likelihood of hallucinated responses. As illustrated in Fig. 1(b), by constructing preferred–rejected response pairs under this controlled setting, we guide the model to overcome its inherent hallucination behaviors and mitigate the generation of erroneous reasoning trajectories.

In summary, our main contributions are as follows:

- We analyze the structural causes of hallucination in multi-image reasoning, identifying imbalanced visual information flow and insufficient cross-image semantic association as key factors that limit multi-image reasoning performance.
- We propose a novel framework CAPL, which integrates selective cross-image attention with preference alignment training, enhancing semantic interaction among critical cross-image tokens and reinforce the model to better perceive and utilize inter-image interactions.
- Extensive experiments(Fig. 1(d)) demonstrate that our method generalizes well across multiple recent vision-language models, significantly reducing hallucination and improving reasoning performance on multi-image tasks.

2 Related Work

2.1 Large Vision-Language Models

Recent years have witnessed significant progress in Large Vision-Language Models across a wide range of multimodal understanding and generation tasks. Representative approaches typically adopt a two-stage framework that aligns a visual encoder with a large language model. For example, the LLaVA series [33] maps visual features into the language space via a projection layer followed by instruction tuning; Qwen-VL [2] and InternVL [6] further enhance high-resolution visual perception and support multi-image inputs; GLM-4V [14] enhances visual understanding while maintaining strong language reasoning capabilities. These methods generally follow a unified architecture of "visual encoder + linear projection + LLM", and rely on instruction-tuning data to achieve multimodal alignment, enabling applications such as multimodal dialogue [36], visual question answering (VQA) [1] and image captioning [39]. However, in multi-image scenarios, existing models typically adopt relatively simple strategies, such as concatenating visual tokens or directly fusing shared representations, without explicitly modeling structured relationships across images. This limitation poses significant challenges for multi-image reasoning and consistency modeling.

2.2 Multimodal Hallucination in LVLMs

Although LVLMs achieve strong performance across various benchmarks, multimodal hallucination remains a persistent challenge, which can be generally categorized to single-image hallucination and multi-image hallucination.

Single-image hallucination. This issue typically stems from data bias [16], annotation noise [30], insufficient vision–language alignment [32, 41], and the model's over-reliance on language priors [13]. To mitigate this problem, existing approaches focus on several directions, including data refinement, bias correction [16], improved cross-modal alignment architectures [18], and the introduction of preference alignment [9,20,52] or decoding constraints [10,24] to improve consistency between generated content and visual evidence. In addition, some studies [11] analyze hallucinations from the perspective of attention distribution and token grounding, revealing that models may attend to irrelevant visual regions during generation. Other works [7] attempt to strengthen visual grounding by incorporating external detectors or region-level supervision. These methods have shown effectiveness in reducing hallucinations in single-image scenarios, but their applicability to more complex multi-image settings remains limited.

Multi-image hallucination. With the emergence of multi-image tasks, recent work has begun to investigate error patterns in multi-image scenarios, such as incorrectly merging information from different images during comparison or introducing nonexistent entities and attributes in relational reasoning. Some training-free methods [26] attempt to alleviate bias toward particular images by averaging attention across images. However, their effectiveness remains limited since the model parameters are not updated. Preference learning approaches [27,37] construct preference pairs based on attention distributions over key images or from global and targeted perspectives. Nevertheless, these works mainly address quasi single-image settings within multi-image inputs, where images are often semantically unrelated, resulting in limited attention to modeling genuine cross-image semantic relationships. Furthermore, the complex dependencies and fine-grained differences between images in multi-image tasks make hallucination problems more difficult to mitigate. Especially when inter-image information flow is inadequately modeled, models are more prone to semantic confusion and erroneous reasoning across images.

3 Method

3.1 Preliminary

In the multi-image setting, we assume that the input consists of N images and their corresponding textual descriptions. The k-th image is encoded by a visual encoder into a sequence of visual tokens:

$$\mathbf{I}_k = [v_{k,1}, \ldots, v_{k,\tau_k}], \quad k = 1, \ldots, N \tag{1}$$

where $v_{k,j}$ denotes the j-th visual token of the k-th image, and τ_k represents the length of the visual token sequence for that image. Similarly, the associated text is tokenized into a sequence of textual tokens:

$$\mathbf{T}_k = [t_{k,1}, \ldots, t_{k,L_k}], \quad k = 1, \ldots, N \tag{2}$$

where $t_{k,j}$ denotes the j-th token of the k-th text segment, and L_k denotes the length of the corresponding text token sequence. Following common LVLM input formatting, the multimodal input is organized according to its original ordering in an interleaved image-text format to form a unified sequence.

$$\mathbf{z} = [\mathbf{I}_1, \mathbf{T}_1, \mathbf{I}_2, \mathbf{T}_2, \ldots, \mathbf{I}_N, \mathbf{T}_N]. \tag{3}$$

Let T denote the total length of the concatenated sequence. Under the autoregressive Transformer framework, self-attention is computed as:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right), \tag{4}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T \times d}$ are the query, key, and value matrices obtained through linear projections, and $\mathbf{M} \in \mathbb{R}^{T \times T}$ is the attention mask matrix. Under the standard causal attention mechanism, $\mathbf{M}$ is defined as:

$$\mathbf{M}_{ij} = \begin{cases} 0, & j \leq i, \\ -\infty, & j > i, \end{cases} \tag{5}$$

which constrains each token to attend only to itself and preceding tokens, thereby ensuring autoregressive generation.

Since both visual and textual tokens are embedded into a unified sequence space, cross-image and cross-modal dependencies are implicitly determined by the relative positions of tokens in the sequence and the causal attention mask.

The overall framework of CAPL is illustrated in Fig. 2.

3.2 Selective Cross-Image Token Interaction

In the multi-image autoregressive modeling framework, the standard causal attention induces a unidirectional cross-image information flow: tokens from later images can attend to representations of earlier images, whereas tokens from earlier images cannot access information from later ones. This asymmetry introduces an explicit ordering bias and weakens the cross-image symmetry required for modeling mutual semantic relationships. To alleviate this limitation, we propose a cross-image attention interaction mechanism that activates attention connections between tokens from different images. By breaking the unidirectional constraint, this design allows earlier images to perceive later ones and enables bidirectional information flow across visual input. We implement this by removing the causal mask between different images while preserving that within each image to retain intra-image positional structure and order bias.

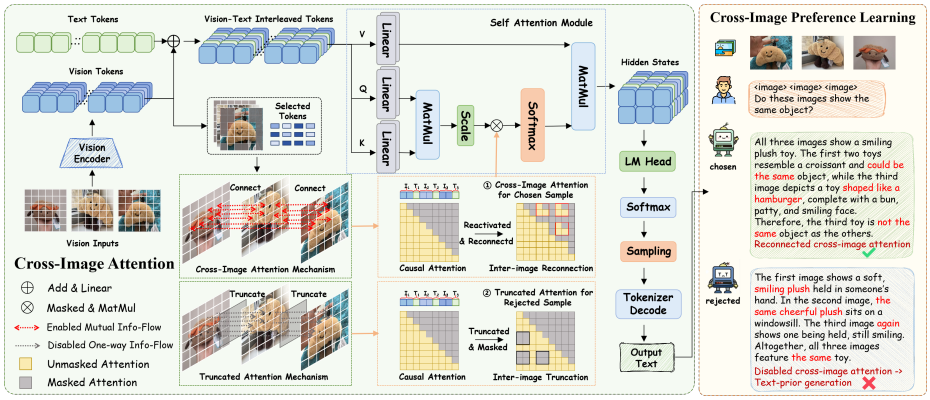

Fig. 2: The overview of the CAPL framework. The pipeline consists of attention modification and preference learning. Image tokens are ranked to select key tokens, and the original causal mask is modified into enhanced cross-image attention (reactivating inter-image token interactions) and truncated cross-image attention (blocking inter-image interactions). The enhanced mask is used for inference and positive sample construction, while the truncated mask generates negative samples for DPO training.

Formally, let $g(i)$ denote the index of the image to which token i belongs. The cross-image mask $\mathbf{M}^{\text{cross}}$ is defined as:

$$\mathbf{M}_{ij}^{\text{cross}} = \begin{cases} 0, & g(i) \neq g(j), \\ \mathbf{M}_{ij}^{\text{causal}}, & g(i) = g(j), \end{cases} \tag{6}$$

which allows bidirectional attention only between tokens from different images, while retaining the original causal attention structure within each image.

Since information density and semantic relevance may vary across images, enabling full cross-image attention may introduce redundant interactions. To address this issue, we further introduce a key-token selection mechanism based on embedding energy. Let the visual tokens of the k-th image be denoted as $\mathbf{h}_{k,1}, \ldots, \mathbf{h}_{k,\tau_k} \in \mathbb{R}^d$, and define their response intensity as:

$$s_{k,i} = \|\mathbf{h}_{k,i}\|_2. \tag{7}$$

We introduce a ratio parameter $\rho \in (0, 1]$ to control the selection range of key tokens. For each image, the top $\lfloor \rho\tau_k \rfloor$ tokens with the highest response intensity are selected to form the key-token set:

$$\mathcal{S}_k = \{i \mid s_{k,i} \text{ ranks among the top } \rho\tau_k\}. \tag{8}$$

Based on the cross-image mask defined in Eq. 6, we further construct the selective cross-image mask $\mathbf{M}^{\text{cross\_sel}}$ as:

$$\mathbf{M}_{ij}^{\text{cross\_sel}} = \begin{cases} 0, & g(i) \neq g(j),\ i \in \mathcal{S}_{g(i)},\ j \in \mathcal{S}_{g(j)}, \\ \mathbf{M}_{ij}^{\text{causal}}, & \text{otherwise}. \end{cases} \tag{9}$$

This mechanism enables effective interactions only between key tokens from different images, thereby enhancing cross-image semantic association modeling while suppressing irrelevant information propagation.

The proposed interaction attention removes the cross-image causal order bias, enabling symmetric visibility across images and facilitating relational reasoning. However, some multi-image tasks involve temporal dependencies or single-image queries, where preserving the sequential structure of images remains necessary. Therefore, completely replacing causal attention with cross-image attention may not be optimal. To balance these requirements, we fuse the selective cross-image attention with the original causal attention and compute the final attention weights as their equal-weight combination:

$$\mathbf{A}^{\text{fuse}} = \frac{1}{2}\left(\mathbf{A}^{\text{causal}} + \mathbf{A}^{\text{cross\_sel}}\right). \tag{10}$$

Furthermore, to prevent cross-image interactions from disrupting the original autoregressive modeling pathway, we adopt an alternating mask strategy at the decoder-layer level: odd-numbered layers apply the selective cross-image mask $\mathbf{M}^{\text{cross\_sel}}$, while even-numbered layers retain the original causal mask $\mathbf{M}^{\text{causal}}$. This hierarchical alternating design enhances cross-image relational modeling while preserving stability and generalization in non-cross-image tasks.

3.3 Cross-Image Attention Guided DPO for Preference Learning

Most vision-language models are pre-trained using an autoregressive causal attention paradigm, so merely modifying the attention mechanism during inference yields only temporary improvements. Traditional SFT can force the model to mimic positive samples but cannot suppress its inherent hallucination tendencies. Existing approaches such as RLHF [41] and RLAIF [23] require costly construction of negative samples and often struggle to capture the model's internal response patterns and latent hallucination directions, thus failing to effectively generate hallucination-targeted negatives. In contrast, Direct Preference Optimization (DPO) offers a more flexible training paradigm. By explicitly contrasting the generation probabilities of positive and negative samples, it guides the model to prefer reliable cross-image outputs while suppressing potential hallucination pathways. This approach not only reshapes the output distribution, but also promotes stable cross-image relational modeling in the parameter space, rather than relying solely on inference-time adjustments.

The construction of positive and negative samples plays a crucial role in the DPO framework. For positive samples, we generate outputs using the selective cross-image attention mechanism and further refine them with Qwen3 to ensure correctness. To effectively guide the model to recognize and correct its hallucination behaviors, we propose to construct negative samples that expose its inherent erroneous reasoning. Directly using outputs from the model as negatives is of limited effectiveness, as these outputs often fail to reveal its hallucination patterns and therefore provide weak optimization signals. Specifically,

we draw inspiration from the causal attention mechanism in multi-image settings from a complementary perspective. While standard causal attention allows information to flow unidirectionally between images, we exploit this property to induce hallucinations by truncating attention between tokens from different images, enforcing representational independence across all images. This forces the model to rely solely on individual images when generating negative samples, while reasoning about cross-image relationships degenerates to inference based on text priors. Formally, we define the truncated attention mask $\mathbf{M}_{\text{trunc}}$ as:

$$\mathbf{M}_{ij}^{\text{trunc}} = \begin{cases} \mathbf{M}_{ij}^{\text{causal}}, & g(i) = g(j), \\ -\infty, & g(i) \neq g(j), \end{cases} \tag{11}$$

where $g(i)$ denotes the index of the image to which token i belongs. Text generated under this mask serves as the negative sample.

Given a $x_{\text{I,T}}$ with both input images and the text question, let y^+ and y^- denote the outputs generated under cross-image attention and the truncated attention mask, respectively. The DPO objective is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\log \sigma\Big(\beta \log \frac{\pi_\theta(y^+ \mid x_{\text{I,T}})}{\pi_{\text{ref}}(y^+ \mid x_{\text{I,T}})} - \beta \log \frac{\pi_\theta(y^- \mid x_{\text{I,T}})}{\pi_{\text{ref}}(y^- \mid x_{\text{I,T}})}\Big), \tag{12}$$

where $\pi_\theta(y \mid x_{\text{I,T}})$ is the probability of generating y given the image–text input $x_{\text{I,T}}$, and $\pi_{\text{ref}}(y \mid x_{\text{I,T}})$ is the reference model probability. Through this training process, the model learns to explicitly prefer outputs produced under cross-image bidirectional attention. As a result, the utilization of cross-image information is gradually strengthened in the parameter space, leading to improved semantic relational modeling across images and effective suppression of multi-image hallucinations. This contrastive training strategy further enables the model to robustly handle complex multi-image inputs while preserving its original language generation capability.

3.4 Training Optimization

While DPO optimizes the relative preference between positive and negative responses, it does not explicitly enforce the model to imitate the token-level generation trajectory of high-quality positive samples. As a result, the model may learn preference ordering without fully internalizing the structured cross-image reasoning process. To address this limitation, we incorporate a negative log-likelihood (NLL) loss on the positive samples. Given input $x_{\text{I,T}}$ and its corresponding positive response $y^+ = \{y_1^+, \ldots, y_T^+\}$, the NLL objective is defined as:

$$\mathcal{L}_{\text{NLL}}(\pi_\theta) = -\sum_{t=1}^{T} \log \pi_\theta\big(y_t^+ \mid x_{\text{I,T}}, y_{<t}^+\big), \tag{13}$$

where π_θ denotes the model policy. This objective encourages the model to follow generation trajectories aligned with cross-image semantic reasoning.

Finally, the overall training objective combines DPO and NLL as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) + \lambda \mathcal{L}_{\text{NLL}}(\pi_\theta), \tag{14}$$

where λ is a balancing coefficient. The DPO term enforces relative preference alignment between contrasted samples, while the NLL term reinforces token-level imitation of reliable cross-image reasoning paths. This hybrid objective improves optimization stability and further enhances cross-image semantic consistency.

4 Experiments

4.1 Experimental Setup

Baselines. We apply our method to three different base models: Qwen2.5-VL-7B-Instruct [3], InternVL2.5-8B [5], and GLM4.1VBase-9B [15]. As baselines, we use the original pre-trained models without additional alignment. For a broader comparison, we report results of other LVLMs, including Idefics2 [22], Idefics3 [21], LLaVA-OV [25], LLaVA-Next [25], InternVL2 [6] and Qwen2VL [45].

Benchmarks. To comprehensively evaluate the effectiveness and generalization ability of our method in multi-image scenarios, we conduct experiments on diverse benchmarks. We first evaluate on BLINK [12] and MUIRBench [43], two representative benchmarks specifically designed to evaluate **multi-image hallucination**, which measure the model's ability to establish semantic associations across images and suppress hallucinated reasoning. We further evaluate on mainstream **multi-image general capability** evaluation benchmarks, including NLVR2 [40], QBench2 [49], MIBench [34], and MIRB [50], which primarily focus on overall general understanding and reasoning rather than explicitly targeting hallucination. In addition, we conduct experiments on **single-image** benchmarks such as POPE [28], CHAIR [39], MMBench [35] and AMBER [44] to examine robustness under cross-domain settings. Through evaluation on these diverse benchmarks, we aim to demonstrate the effectiveness and robustness of our approach across different task types and domain scenarios.

Implementation Details We construct a training set of 3.6K samples collected from WikiArt, MDVP [29], Mantis-Instruct [19], and VLM2Bench [47], ensuring that no data overlaps with the evaluation benchmarks. Considering the architectural differences among base models, we adopt model-specific hyperparameters. For the top cross-image token selection ratio ρ, the values for Qwen2.5-VL, InternVL2.5, and GLM4.1VBase are set to 0.95, 0.95, and 0.9, respectively. For the NLL loss weight λ, the corresponding values are 2.5, 2.5, and 2. During training, we use a learning rate of $1e{-}4$ and apply LoRA fine-tuning with a rank of 16. Regarding training frameworks, Qwen2.5-VL and GLM4.1VBase are trained with LLaMAFactory [53], while InternVL2.5 is trained with MS-Swift [51]. All models are evaluated on VLMEvalKit [8] to ensure consistency and comparability, and all experiments are conducted on NVIDIA L20 GPUs.

Table 1: Performance comparison on hallucination and general capability evaluation in multi-image scenarios. BLINK and MUIRBench are hallucination tasks, which constitute the core evaluation of our method. "–" indicates results that cannot be measured.

Models	Params	Hallucination Eval		General Capability Eval			
		BLINK	MUIRB	NLVR2	QBench2	MIBench	MIRB
Idefics2	8B	45.24	29.85	56.81	38.6	49.28	37.56
Idefics3	8B	48.34	30.96	85.14	64.2	54.18	–
LLaVA-OV	7B	44.77	30.85	86.82	70.1	62.37	47.3
LLaVA-Next	7B	41.19	30.50	50.34	39.5	–	–
InternVL2	7B	50.34	45.61	77.68	69.8	59.37	53.15
Qwen2VL	7B	53.17	39.57	87.41	76.8	69.20	31.68
Qwen2.5-VL	7B	54.60	58.42	79.85	71.1	72.42	52.73
+CAPL (Ours)	7B	**57.76**	**62.00**	80.05	72.4	71.06	56.55
InternVL2.5	8B	54.81	48.54	90.42	75.5	63.42	54.18
+CAPL (Ours)	8B	**55.76**	**52.12**	90.13	75.3	65.11	56.55
GLM4.1VBase	9B	58.17	57.84	84.98	74.4	70.86	59.96
+CAPL (Ours)	9B	**61.33**	**60.57**	84.87	73.6	71.70	60.06

4.2 Main Results

Results on Multi-Image Hallucination Tasks. Results on the multi-image hallucination benchmarks BLINK and MUIRBench are shown in Tab. 1. Overall, CAPL brings consistent improvements across Qwen2.5-VL, InternVL2.5, and GLM4.1VBase, demonstrating strong cross-architecture generalization. On BLINK, all models achieve gains of approximately 1–3 points. On the large-scale MUIRBench benchmark, which emphasizes complex cross-image relational reasoning, the improvements are more pronounced, reaching over 3.5 points in the best case. Notably, even for the strong baseline GLM4.1VBase-9B, incorporating CAPL still yields steady improvements. This observation suggests that even advanced models rely on relatively naive attention mechanisms in multi-image settings, leaving room for improving inter-image interaction structures. Overall, explicitly modeling cross-image attention together with preference optimization consistently enhances the reliability of multi-image reasoning and effectively mitigates hallucinations caused by weak cross-image associations.

To further validate the effectiveness of CAPL, we additionally compare it with recent representative multi-image alignment methods, SOFA and MIA-DPO, under the same model architectures, as shown in Table 2. Across different models and benchmarks, SOFA yields only limited improvements over the base models, with relatively small overall gains. MIA-DPO achieves certain improvements in some settings, but its performance varies across different models and tasks. In contrast, CAPL consistently achieves the best performance in all experimental settings, significantly outperforming both methods on the BLINK and MUIRBench benchmarks. These results indicate that, in multi-image scenarios,

Table 2: Performance comparison with SOFA and MIA-DPO.

Model	Benchmark	Base	SOFA	MIA	CAPL
Qwen2.5-VL	BLINK	54.60	54.92	56.50	**57.76**
	MUIRBench	58.42	59.34	56.15	**62.00**
InternVL2.5	BLINK	54.81	55.02	54.45	**55.76**
	MUIRBench	48.54	49.07	47.88	**52.12**
GLM4.1V	BLINK	58.18	58.18	59.76	**61.34**
	MUIRBench	57.84	57.23	57.23	**60.27**

Table 3: Performance comparison on single-image benchmarks. ↑ indicates higher is better; ↓ indicates lower is better.

Models	POPE ↑	CHAIRi ↓	CHAIRs ↓	MMB ↑	AMBER ↑
Idefics2	86.35	4.8	7.6	76.46	86.40
Idefics3	85.50	7.6	40.2	75.51	84.09
LLaVA-OV	89.15	8.7	34.0	81.70	83.99
LLaVA-Next	87.42	13.3	22.6	69.93	84.96
InternVL2	84.39	8.3	34.0	82.90	86.00
Qwen2VL	84.00	10.5	28.8	81.53	85.96
Qwen2.5-VL	81.23	**7.2**	29.6	**88.15**	85.13
+CAPL (Ours)	**82.94**	7.5	**28.6**	87.48	**85.25**
InternVL2.5	88.94	**6.4**	24.4	84.11	88.99
+CAPL (Ours)	**89.08**	7.3	**23.8**	**84.27**	**89.79**
GLM4.1VBase	84.79	7.2	22.0	84.36	**89.20**
+CAPL (Ours)	**86.20**	**6.5**	**18.4**	**84.62**	88.49

existing training-free or weak alignment strategies are still insufficient to fully model cross-image dependencies. In comparison, explicitly modeling cross-image attention together with preference-based optimization more effectively improves the reliability and stability of multi-image reasoning.

Results on Multi-Image General Tasks. We further assess CAPL on general multi-image general benchmarks in Tab. 1, including NLVR2, QBench2, MIBench, and MIRB. These tasks evaluate whether the model can determine the correctness of descriptions given multiple images, emphasizing multi-image reasoning and general capability rather than hallucination detection. Although CAPL is primarily designed to mitigate hallucinations in multi-image tasks, we observe that its performance remains stable or even shows slight improvements across most multi-image general benchmarks. We attribute this gain to explicit cross-image attention modeling, which regulates information flow between images and increases the effective contribution of visual tokens during reasoning, thereby strengthening the model's reliance on visual evidence.

Table 4: Ablation study on our framework components: we evaluate results by progressively adding our cross-image attention and attentive preference training.

	Qwen2.5-VL		InternVL2.5		GLM4.1VBase	
	BLINK	MUIRBench	BLINK	MUIRBench	BLINK	MUIRBench
Base	54.60	58.42	54.81	48.54	58.17	57.84
+Attn	55.34	58.96	55.02	49.07	58.23	58.07
Ours	57.76	62.00	55.76	52.12	61.33	60.57

Results on Single-Image Tasks. As shown in the Tab. 3, although CAPL is primarily trained on multi-image samples and does not directly involve single-image task data, it still maintains stable performance across several single-image visual benchmarks, with improvements on some of them. For example, on Qwen2.5-VL, the POPE score increases from 81.23 to 82.94, while CHAIRs decreases from 22 to 18.4 on GLM4.1VBase. We conjecture that during training, CAPL effectively suppresses the model's latent hallucination tendencies by pushing it away from negative samples, while preference learning over multi-image visual information enriches the model's visual knowledge. As a result, the model can use visual signals more accurately even when reasoning over a single image.

Table 5: Ablation on negative sample construction across all subsets of MUIRBench. We compare the Base Model of GLM4.1VBase, DPO with original negatives, and DPO with truncated negatives.

Model	Overall	Action	Similarity	Cartoon	Counting	Diagram	Difference	Geographic	I-T Match	Ordering	Scene	Grounding	Retrieval
GLM4.1VBase	57.9	46.3	58.7	46.2	39.3	73.9	57.9	41.0	68.8	20.3	66.1	28.6	59.6
+Attn	58.1	47.0	56.6	46.2	40.2	74.9	57.4	41.0	68.8	23.4	65.6	28.6	61.0
⊕ Original DPO	59.5	46.3	62.8	47.4	40.6	75.6	58.5	44.0	73.3	20.3	69.4	31.0	55.8
⊕ Truncated DPO	60.6	48.2	59.7	48.7	42.7	76.9	60.3	48.0	72.6	21.9	70.4	31.0	59.6
Δ	+2.7	+1.8	+1.0	+2.6	+3.4	+3.0	+2.4	+7.0	+3.9	+1.6	+4.3	+2.4	+0.0

4.3 Ablation Studies

The Effect of Cross-Image Attention and Training Framework. From Tab. 4, we observe that introducing selective cross-image attention (+Attn) consistently yields modest yet stable improvements across all three backbone models on both BLINK and MUIRBench. This indicates that enhancing cross-image in-

formation interaction at inference time alone can partially alleviate erroneous associations in cross-image reasoning, although the overall gains remain limited.

When further combined with our proposed preference optimization training, the performance improvements become substantially more pronounced, particularly on MUIRBench (*e.g.*, Qwen2.5-VL improves to 62.00 and GLM4.1VBase reaches 60.57). This observation suggests a synergistic effect between structured cross-image modeling and targeted preference optimization: the former explicitly strengthens the model's ability to capture inter-image dependencies, while the latter leverages preference signals to further regularize the generation process, thereby more effectively mitigating multi-image hallucination.

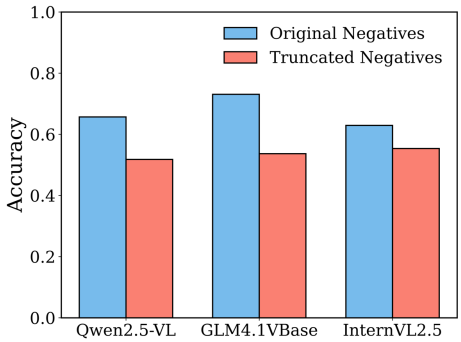

Fig. 3: Comparison of three models' accuracies on two types of negative samples: those generated by the original model structure (Original Negatives) and those generated using the truncated attention structure (Truncated Negatives). Lower accuracy indicates that the negative samples are more challenging.

The Effect of Truncated Attention–Induced Negatives. In DPO training, the quality of negative samples is crucial, as samples that deviate further from the true distribution typically provide stronger optimization signals. To verify the effectiveness of Truncated Attention–Induced Negatives, we first compared the accuracy of negative samples generated by the original model with those generated using truncated attention. The results in Fig. 3 show that the induced negatives exhibit consistently lower accuracy, decreasing by about 20% on GLM4.1VBase. This result indicates that truncating cross-image attention effectively exposes more challenging error cases. Building on this observation, we conducted comparative experiments across all MUIRBench subtasks using GLM4.1VBase, including the Base model, +Attention, original DPO, and Truncated DPO. The results in Tab. 5 demonstrate that Truncated DPO performs particularly well on representative tasks, such as Geographic Understanding (+7.0), Scene Understanding (+4.3), and Difference Spotting (+2.4), outperforming the original DPO on most subtasks. By restricting cross-image information flow during negative sample generation, truncated attention constructs more challenging erroneous samples, resulting in a clearer preference gap during DPO training. Additionally, we designed some negative samples whose final answers are correct, but due to blocked inter-image attention, the reasoning paths are incorrect; these are also used in DPO to help the model learn to avoid potential errors.

Ablation on the Select Ratio ρ. The selection ratio ρ is a key parameter in cross-image attention, controlling the number of tokens involved in cross-image interactions. When ρ is too small, information flow is restricted and cross-image alignment becomes weaker. conversely, an excessively large ρ introduces

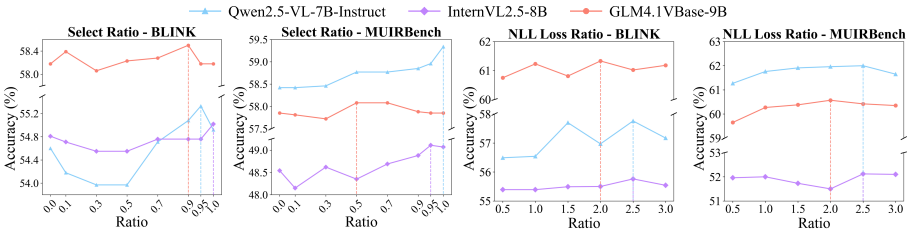

Fig. 4: Ablation results with respect to the Select Ratio ρ and the NLL Loss Ratio λ.

redundant noise and amplifies interference. The results in Fig. 4(Left) show that performance improves as ρ increases but declines slightly when it approaches 1, indicating that a small portion of image tokens may contain noise. Retaining most, but not all, key tokens for cross-image interaction achieves a better balance between effective modeling and noise suppression. Empirically, the optimal values are $\rho = 0.95$ for Qwen2.5-VL and InternVL2.5, and $\rho = 0.9$ for GLM4.1VBase.

Ablation on the NLL Loss Ratio λ. The NLL loss ratio λ determines the weight of the NLL loss in the overall training objective. When λ is too small, the model relies excessively on the DPO signal, which may weaken its language modeling capability, while a overly large λ diminishes the preference distinction signal and may affect optimization stability. Experimental results in Fig. 4(Right) show that, considering performance on multi-image hallucination benchmarks BLINK and MUIRBench, Qwen2.5-VL and InternVL2.5 achieve optimal performance at $\lambda = 2.5$, whereas GLM4.1VBase performs best at $\lambda = 2$.

5 Conclusion

This work proposes **C**ross-Image **A**ttention calibration and **P**reference **L**earning **(CAPL)**, a framework designed for multi-image tasks. It enhances inter-image information flow through selective cross-image attention and introduces a preference optimization strategy based on a contrast of reasoning results between full cross-image interaction and truncated images, encouraging predictions grounded in true visual evidence. Experimental results show that each component of our method contributes to performance improvements, and the combined framework significantly enhances model performance on multi-image hallucination and general tasks, while maintaining single-image reasoning capabilities, demonstrating the practicality and effectiveness of cross-image attention and attentive preference learning design in complex multi-image scenarios.

6 Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under grant 62301189, 62576122,62571298, Guangdong Basic and Applied Basic Research Foundation under grant 2026A1515011139.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bai, T., Fan, Y., Qiu, J., Sun, F., Song, J., Han, J., Liu, Z., He, C., Zhang, W., Yuan, B.: Hallucination at a glance: Controlled visual edits and fine-grained multimodal learning. arXiv preprint arXiv:2506.07227 (2025)
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
7. Dai, W., Liu, Z., Ji, Z., Su, D., Fung, P.: Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2136–2148 (2023)
8. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 11198–11201 (2024)
9. Fang, H., Li, J., Kong, J., Zhuang, T., Gao, K., Chen, B., Xia, S.T., Wang, Y.: Seeing through the chain: Mitigate hallucination in multimodal reasoning models via cot compression and contrastive preference optimization. arXiv preprint arXiv:2602.03380 (2026)
10. Fang, H., Zhou, C., Kong, J., Gao, K., Chen, B., Xia, S.T.: Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. arXiv preprint arXiv:2505.19678 (2025)
11. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14303–14312 (2024)
12. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
13. Ghosh, S., Evuru, C.K.R., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., Manocha, D.: Visual description grounding reduces hallucinations and boosts reasoning in lvlms. arXiv preprint arXiv:2405.15683 (2024)
14. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500 (2024)

15. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv e-prints pp. arXiv–2507 (2025)
16. Hu, H., Zhang, J., Zhao, M., Sun, Z.: Ciem: Contrastive instruction evaluation method for better instruction tuning. arXiv preprint arXiv:2309.02301 (2023)
17. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
18. 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)
19. Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., Chen, W.: Mantis: Interleaved multi-image instruction tuning. arXiv preprint arXiv:2405.01483 (2024)
20. Kong, J., Fang, H., Liao, S., Li, J., Chen, B., Wu, H., Xia, S.T., Zhang, M.: Reasoning matters: Mitigate hallucination in multimodal large reasoning models via reasoning-conditioned preference optimization. arXiv preprint arXiv:2605.27906 (2026)
21. Laurençon, H.: Foundation Vision-Language models. Ph.D. thesis, Sorbonne Université (2025)
22. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? Advances in Neural Information Processing Systems **37**, 87874–87907 (2024)
23. Lee, H., Phatale, S., Mansoor, H., Lu, K.R., Mesnard, T., Ferret, J., Bishop, C., Hall, E., Carbune, V., Rastogi, A.: Rlaif: Scaling reinforcement learning from human feedback with ai feedback (2023)
24. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: 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. pp. 13872–13882 (2024)
25. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
26. Li, J., Wu, M., Jin, Z., Chen, H., Ji, J., Sun, X., Cao, L., Ji, R.: Mihbench: Benchmarking and mitigating multi-image hallucinations in multimodal large language models. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3143–3152 (2025)
27. Li, X., Zhang, M., Chen, P., Zheng, X., Zhang, Y., Zheng, J., Shen, Y., Li, K., Fu, C., Sun, X., et al.: Zooming from context to cue: Hierarchical preference optimization for multi-image mllms. arXiv preprint arXiv:2505.22396 (2025)
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
29. Lin, W., Wei, X., An, R., Gao, P., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. arXiv preprint arXiv:2403.20271 (2024)
30. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)

31. 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)
32. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
34. Liu, H., Zhang, X., Xu, H., Shi, Y., Jiang, C., Yan, M., Zhang, J., Huang, F., Yuan, C., Li, B., et al.: Mibench: Evaluating multimodal large language models over multiple images. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 22417–22428 (2024)
35. 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)
36. Liu, Z., Chu, T., Zang, Y., Wei, X., Dong, X., Zhang, P., Liang, Z., Xiong, Y., Qiao, Y., Lin, D., et al.: Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. Advances in Neural Information Processing Systems **37**, 8698–8733 (2024)
37. Liu, Z., Zang, Y., Dong, X., Zhang, P., Cao, Y., Duan, H., He, C., Xiong, Y., Lin, D., Wang, J.: Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. arXiv preprint arXiv:2410.17637 (2024)
38. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
39. 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. pp. 4035–4045 (2018)
40. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 6418–6428 (2019)
41. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 13088–13110 (2024)
42. Tian, X., Zou, S., Yang, Z., Zhang, J.: Identifying and mitigating position bias of multi-image vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10599–10609 (2025)
43. Wang, F., Fu, X., Huang, J.Y., Li, Z., Liu, Q., Liu, X., Ma, M.D., Xu, N., Zhou, W., Zhang, K., et al.: Muirbench: A comprehensive benchmark for robust multi-image understanding. arXiv preprint arXiv:2406.09411 (2024)
44. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397 (2023)
45. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
46. Yang, X., Zhang, H., Qi, G., Cai, J.: Causal attention for vision-language tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9847–9857 (2021)

47. Zhang, J., Yao, D., Pi, R., Liang, P.P., Fung, Y.R.: Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7510–7545 (2025)
48. Zhang, Y., Ding, Y., Zhang, S., Zhang, X., Li, H., Li, Z.z., Wang, P., Wu, J., Ji, L., Shen, Y., et al.: Perl: Permutation-enhanced reinforcement learning for interleaved vision-language reasoning. arXiv preprint arXiv:2506.14907 (2025)
49. Zhang, Z., Wu, H., Zhang, E., Zhai, G., Lin, W.: Q-bench ++: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10404–10418 (2024)
50. Zhao, B., Zong, Y., Zhang, L., Hospedales, T.: Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. arXiv preprint arXiv:2406.12742 (2024)
51. Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., et al.: Swift: a scalable lightweight infrastructure for fine-tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 29733–29735 (2025)
52. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023)
53. Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z.: Llamafactory: Unified efficient fine-tuning of 100+ language models. In: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations). pp. 400–410 (2024)