Zooming from Context to Cue: Hierarchical Preference Optimization for Multi-Image MLLMs

Xudong Li[1*] **Mengdan Zhang**[2*] **Peixian Chen**[2] **Xiawu Zheng**[1] **Yan Zhang**[1†]
Jingyuan Zheng[1] **Yunhang Shen**[2] **Ke Li**[2] **Chaoyou Fu**[3] **Xing Sun**[2] **Rongrong Ji**[1]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China
[2] Tencent Youtu Lab [3] Nanjing University
{lxd761050753, zhangmengdanrz}@gmail.com, {zhengxiawu, rrji}@xmu.edu.cn

Abstract

Multi-modal Large Language Models (MLLMs) excel at single-image tasks but struggle with multi-image understanding due to cross-modal misalignment, leading to hallucinations (context omission, conflation, and misinterpretation). Existing methods using Direct Preference Optimization (DPO) constrain optimization to a solitary image reference within the input sequence, neglecting holistic context modeling. To address this, we propose **C**ontext-to-**C**ue **D**irect **P**reference **O**ptimization **(CcDPO)**, a multi-level preference optimization framework that enhances per-image perception in multi-image settings by zooming into visual clues—from sequential context to local details. Our approach features two sequentially dependent components: (i) *Context-Level Optimization:* By introducing low-cost sequence preference pairs, we optimize the model to distinguish between complete and disrupted multi-image contexts, thereby correcting cognitive biases in MLLMs' multi-image understanding. (ii) *Needle-Level Optimization:* By integrating region-specific visual prompts with multimodal preference supervision, we direct the model's attention to critical visual details, effectively suppressing perceptual biases toward fine-grained visual information. To support scalable optimization, we also construct **MultiScope-42k**, an automatically generated multi-image dataset with hierarchical preference pairs. Experiments show that CcDPO significantly reduces hallucinations and yields consistent performance gains across general single- and multi-image tasks. Codes are available at https://github.com/LXDxmu/CcDPO.

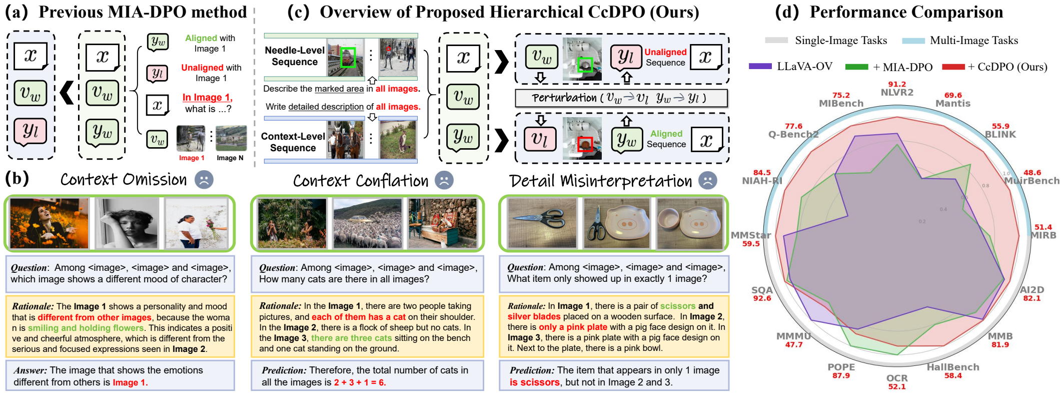
1 Introduction

Simultaneously understanding multiple images remains a fundamental yet underexplored challenge for Multi-modal Large Language Models (MLLMs) [1, 2, 3, 4]. Despite MLLMs excelling in single-image tasks like visual question answering (VQA) [5, 6, 7], code generation [8, 9], and storytelling [10, 11], and open-source models such as LLaVA [12], BLIP-2 [13], and InternVL [14] showing competitive results on benchmarks including VQAv2 [15], OKVQA [16], and MMMU [17], their capabilities in multi-image contexts are notably constrained. These models frequently struggle with tasks demanding cross-image comparison, spatial reasoning, or temporal alignment [18], often resulting in hallucinations like context omission, conflation, and misinterpretation of local details. These deficiencies ultimately compromise model reliability. The root cause lies in the weak cross-modal alignment within MLLMs, which frequently fails to integrate visual and textual information

---

[*] Equal Contribution.
[†] Corresponding Author.

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



Question: Among <image>, <image> and <image>, which image shows a different mood of character?

Rationale: The **Image 1** shows a personality and mood that is different from other images, because the woma n is smiling and holding flowers. This indicates a positi ve and cheerful atmosphere, which is different from the serious and focused expressions seen in **Image 2**.

Answer: The image that shows the emotions different from others is Image 1.

Question: Among <image>, <image> and <image>, How many cats are there in all images?

Rationale: In the **Image 1**, there are two people taking pictures, and each of them has a cat on their shoulder. In the **Image 2**, there is a flock of sheep but no cats. In the **Image 3**, there are three cats sitting on the bench and one cat standing on the ground.

Prediction: Therefore, the total number of cats in all the images is **2 + 3 + 1 = 6.**

Question: Among <image>, <image> and <image>, What item only showed up in exactly 1 image?

Rationale: In **Image 1**, there is a pair of **scissors and silver blades** placed on a wooden surface. In **Image 2**, there is **only a pink plate** with a pig face design on it. In **Image 3**, there is a pink plate with a pig face design on it. Next to the plate, there is a pink bowl.

Prediction: The item that appears in only 1 image **is scissors**, but not in Image 2 and 3.

Figure 1: (a) Prior multi-image DPO (e.g., MIA-DPO) is constrained by its reliance on predefined image references and text-only preferences, limiting holistic context modeling. (b) These limitations commonly lead to failures such as Context Omission (ignoring relevant images), Context Conflation (misattributing content across images), and Detail Misinterpretation (misrepresenting fine-grained visual cues). (c) CcDPO addresses these issues by hierarchically enhancing MLLMs' visual perception, from overall multi-image contexts to specific fine-grained details. (d) Benchmark comparisons demonstrate CcDPO's improved reasoning capabilities on both multi-image and single-image tasks.

coherently and comprehensively. This limitation becomes particularly pronounced in multi-image settings, where accurate reasoning requires both: (i) Precise interpretation of intra-image regional details, and (ii) establishing meaningful inter-image connections through contextual integration.

To overcome these limitations, instruction tuning with multi-image supervision has been adopted in recent models such as Flamingo [19], IDEFICS [20], and Emu2 [21]. However, these approaches rely on large-scale annotated data, which is costly to construct due to the complexity of modeling inter-image relationships. As a lightweight alternative, *Direct Preference Optimization* (DPO) [22] has emerged as a promising training paradigm, aligning model outputs with human preferences through pairwise supervision without requiring large-scale labeled data, significantly reducing reliance on costly annotations. Recent work has extended DPO to multimodal tasks [23, 24, 25, 26], with MIA-DPO [27] specifically pioneering its application to multi-image scenarios. As shown in Fig. 1(a), MIA-DPO conditions responses on a specific image through explicit query references (e.g., "In Image 1, what is...?"), helping the model associate questions with the correct visual input in context.

While this anchoring strategy mitigates referential ambiguity in MLLMs, the absence of explicit contextual modeling in its architectural design fundamentally limits comprehensive cross-image integration. This constraint is further exacerbated by the high vulnerability of inter-image attention mechanisms to multi-image interference [28, 29, 30]. Consequently, without explicit image references, such approaches often demonstrate limited capacity in autonomously capturing sequential visual context and fine-grained details, giving rise to multiple forms of multi-image hallucinations: **Context Omission**: The model selectively ignores subsets of input images, generating responses based on incomplete sequences (e.g., ignoring Image 3; Fig. 1(b), left). **Context Conflation**: The model erroneously attributes visual elements across images (e.g., describing a cat from Image 3 as appearing in Image 1; Fig. 1(b), middle). **Detail Misinterpretation**: Critical visual details in a certain image are either missed or misinterpreted (e.g., without explicit image-specific instructions, the model fails to recognize the scissors in Image 2 and erroneously detects silver blades in Image 1; Fig. 1(b), right).

To address these challenges, we propose **C**ontext-to-**C**ue **D**irect **P**reference **O**ptimization **(CcDPO)**, a *two-level preference optimization* framework that enhances MLLMs' capability to accurately perceive visual information across hierarchical levels—from sequential multi-image contexts to individual fine-grained details (as shown in Fig. 1(c)). Specifically, it consists of two levels of alignment:

(1) Context-Level Optimization: We formulate structured multi-image captioning as a proxy task to enforce global contextual alignment. By explicitly decomposing model responses into per-image descriptions (e.g., "For Image 1: <caption 1>", "For Image 2: <caption 2>"), we compel the model to accurately attend to each image within the sequence. This structured formulation ensures contextual completeness while mitigating inter-image interference. To further address **Context Omission** and **Context Conflation**, we introduce two perturbation techniques—*sequence truncation* and *content* *swapping*—into the captioning preference optimization process. By training the model to distinguish coherent contexts from disrupted ones, we promote holistic reasoning across the entire input sequence.

(2) Needle-Level Optimization: To address **Detail Misinterpretation**, we propose a fine-grained preference learning strategy that sharpens the model's sensitivity to critical visual cues. Our approach integrates region-focused visual prompts into the preference data and employs DPO training to bias the model toward descriptions aligned with highlighted regions. This enhances the model's ability to detect, attend to, and describe salient visual elements across multiple images. Furthermore, inspired by [31], we incorporate vision contrastive preference supervision by constructing image pairs with varying alignment to reference descriptions. This encourages the model to refine its preference judgments on fine-grained visual cues within each image under contextual settings.

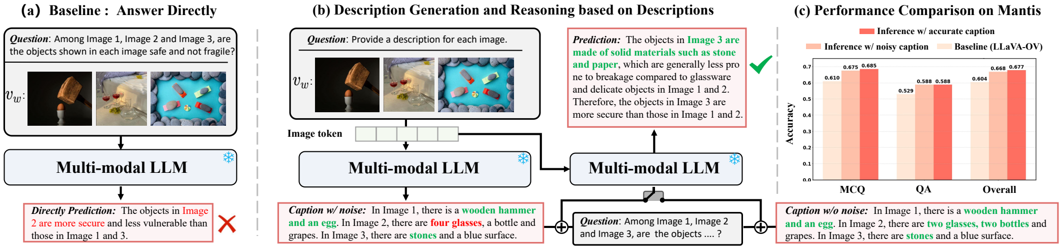
To support these two-level optimization objectives, we introduce **MultiScope-42k**, a scalable multi-image preference dataset. The dataset comprises high-quality *chosen* responses—synthesized by splicing together accurate image- and region-level descriptions—alongside *rejected* responses generated through targeted perturbations at both contextual and local detail levels. This pipeline converts abundant single-image data into complex multi-image training signals, providing a scalable solution to data scarcity. Our main contributions are summarized as follows:

- We pioneer the investigation of cognitive bias in multi-image comprehension for MLLMs, categorizing three prevalent hallucination types. To address these challenges, we propose **C**ontext-to-**C**ue **D**irect **P**reference **O**ptimization **(CcDPO)**, an innovative two-level preference optimization framework that enhances per-image perception in multi-image settings by analyzing visual clues—from sequential context to local details.
- We design a low-cost Context-Level Optimization mechanism, incorporating structured multi-image captioning preferences and targeted perturbation techniques to ensure MLLMs' comprehensive and consistent global context understanding. Complementarily, we develop a Needle-Level Optimization mechanism that enhances fine-grained visual acuity through the integration of region-focused visual prompts and vision contrastive preference signals.
- We construct **MultiScope-42k**, a large-scale, high-quality dataset for two-level multi-image preference learning. The fully automatic generation pipeline is cost-effective and scalable across diverse data sources. After direct preference optimization on this dataset, our method significantly reduces hallucinations and achieves superior performance on multi-image tasks.

2 Related Work

Multi-modal Large Language Models. Recent advances in MLLMs [12, 32, 33, 34] have combined powerful large language models (LLMs) with visual encoders via lightweight connectors, achieving impressive performance across dialogue [35], visual question answering (VQA) [5], and image captioning tasks [36]. These models are typically trained on image-text pairs with instruction tuning, yielding strong single-image understanding. However, they remain prone to hallucinations [37, 38, 39], especially in multi-image scenarios where accurate reasoning requires modeling not only individual images but also their cross-image relationships. Recent studies aim to advance multi-image understanding by incorporating image-text interleaved data [40, 41] during model training. This approach helps develop capabilities such as image comparison [42, 43], cross-image association [44, 45], and temporal reasoning [46, 47]. Nevertheless, instruction tuning with such data remains costly due to the need for complex, fine-grained annotations—an issue exacerbated in multi-image settings.

Direct Preference Optimization. Reinforcement Learning from Human Feedback (RLHF) [48, 39, 3] aligns LLMs with human preferences by training a reward model to maximize the gap between chosen and rejected responses. As a more efficient alternative, Direct Preference Optimization (DPO) [22] bypasses reward modeling and directly optimizes on preference pairs. Recent work investigates DPO's generalization and stability across tasks [26, 49, 50, 51], with multimodal extensions [31, 52, 53, 26] to reduce hallucinations and enhance vision-language grounding in single-image settings. However, current language-based DPO methods often neglect visual details. To address this, vision contrastive DPO approaches either disrupt images [23, 52] or highlight key visual tokens [54, 55], enhancing preference learning but focusing mainly on single-image tasks. MIA-DPO [27] pioneered DPO to multi-image settings by anchoring prompts to specific images, achieving promising results on relevant benchmarks. However, its reliance on predefined references limits holistic context modeling and autonomous cross-image reasoning. In contrast, we propose CcDPO, which explicitly models global context and fine-grained visual cues to enhance multi-image reasoning.


(a) Baseline : Answer Directly
v_w:
Multi-modal LLM
(b) Description Generation and Reasoning based on Descriptions
v_w:
Image token
Multi-modal LLM
Multi-modal LLM
(c) Performance Comparison on Mantis
Inference w/ accurate caption
Inference w/ noisy caption
Baseline (LLaVA-OV)
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.610
0.675
0.685
0.529
0.588
0.588
0.604
0.668
0.677
MCQ
QA
Overall

Question: Among Image 1, Image 2 and Image 3, are the objects shown in each image safe and not fragile?

Directly Prediction: The objects in Image 2 are more secure and less vulnerable than those in Image 1 and 3.

Question: Provide a description for each image.

Caption w/ noise: In Image 1, there is a wooden hammer and an egg. In Image 2, there are four glasses, a bottle and grapes. In Image 3, there are stones and a blue surface.

Prediction: The objects in Image 3 are made of solid materials such as stone and paper, which are generally less pro ne to breakage compared to glassware and delicate objects in Image 1 and 2. Therefore, the objects in Image 3 are more secure than those in Image 1 and 2.

Question: Among Image 1, Image 2 and Image 3, are the objects .... ?

Caption w/o noise: In Image 1, there is a wooden hammer and an egg. In Image 2, there are two glasses, two bottles and grapes. In Image 3, there are stones and a blue surface.

Figure 2: (a) Baseline: Direct inference without context as a condition. (b) Two-stage approach: Generating image captions, then reasoning over them. (c) Performance: Accurate caption understanding as context substantially improves VQA accuracy, with noisy captions also proving beneficial. This highlights deficient intrinsic captioning in MLLMs as a key bottleneck, motivating its enhancement.

Visual Prompting for MLLMs. Visual prompting has been widely used in vision models [56, 57, 58]. Manually annotated points, boxes, or masks—often encoded by a separate prompt encoder—can guide the model to adjust segmentation granularity or select specific instances. More recently, MLLMs have shown the ability to interpret visual prompts directly embedded in the image without additional prompt encoders [59, 60]. Unlike prior work, our method actively integrates visual prompts into preference data and uses DPO training to encourage the model to prefer descriptions aligned with prompted regions, which enhances the model's sensitivity to visually grounded information.

3 Exploring Cognitive Bias in Multi-Image Comprehension of MLLMs

In this section, we investigate the cognitive biases of MLLMs in multi-image comprehension, identifying that deficient internal multi-image captioning ability is a core bottleneck for complex reasoning. Then, we introduce a systematic evaluation methodology to diagnose and quantify three fundamental multi-image hallucination types. Our results confirm that even state-of-the-art models exhibit severe performance degradation and elevated hallucination rates as input images increase.

Limited Contextual Awareness. We identify a fundamental limitation in current MLLMs: their impaired ability to perceive and integrate partial context information for coherent multi-image understanding severely degrades performance. To investigate this limitation, we systematically examine how context quality affects multi-image understanding through controlled experiments with LLaVA-OV-7B [32]. As illustrated in Fig. 2, we evaluate the model's reasoning capability under three distinct context conditions: (1) Accurate Context: The model received the images along with accurate, descriptive captions. (2) Noisy Context: It received the images along with flawed, machine-generated captions. (3) Baseline (No Explicit Context): It received only the images, forcing it to rely entirely on its own internal understanding. As shown in Fig. 2 (c), while accurate captions helped slightly more than noisy ones, the most dramatic finding was a massive **7.37-point accuracy drop** when the model performed inference directly from the images, without any context augmentation. This gap reveals the model's heavy reliance on external caption aids, confirming that its intrinsic captioning capability is a core bottleneck for multi-image reasoning and motivating our direct efforts to enhance this skill.

Multi-image Captioning Re-evaluation. We design a caption generation task as a proxy to systematically evaluate MLLMs' multi-image understanding, identifying three fundamental hallucination types—*Context Omission*, *Context Conflation*, and *Detail Misinterpretation*—that critically degrade model performance. To enable this evaluation, we construct **Context-AMBER-1K** by systematically concatenating images from the single-image AMBER dataset [37] into sequences of two types: short-context sequences with 4 images and long-context sequences with 8 images. Each input is paired with the prompt: ‘‘`Please sequentially describe each of the images shown above. Use the following format: For Image *:⟨description⟩.`" The expected output format is: ‘‘`For Image 1:⟨caption 1⟩,For Image 2:⟨caption 2⟩,..., For Image N:⟨caption N⟩.`", ensuring comprehensive coverage of all input images. This controlled format allows precise identification of hallucination behaviors during multi-image comprehension.

We assess caption quality through four complementary metrics: (a) CHAIR [36] measures object hallucination severity, (b) Response-level Hallucination Rate (Hal) [37] quantifies incorrect descriptions, (c) Cognition-based Hallucination (Cog) [37] detects reasoning errors, and (d) Sequence Coverage Rate (SCover) evaluates caption completeness across image sequences. Notably, for each input sequence, we compute hallucination scores (a), (b), and (c) for each image independently, then average these to get the sequence-level hallucination scores.

Table 1: **Hallucination and preference alignment results.** We report metrics on our constructed multi-image AMBER benchmark. Lower scores indicate better performance for CHAIR, Hal, and Cog, while higher is better for SCover. CcDPO achieves consistent improvements under both 4-image and 8-image settings, effectively reducing hallucinations in contextual multi-image understanding.

		Context-AMBER (4 Images)				**Context-AMBER (8 Images)**			
Models	**Parameter**	CHAIR$\downarrow$	SCover$\uparrow$	Hal$\downarrow$	Cog$\downarrow$	CHAIR$\downarrow$	SCover$\uparrow$	Hal$\downarrow$	Cog$\downarrow$
LLaVA-OV	7B	10.2	74.0%	31.8	2.6	50.6	10.3%	69.1	6.5
+ MIA-DPO [27]	7B	8.9	83.9%	29.8	2.1	28.2	36.7%	45.0	3.8
+ CcDPO (Ours)	7B	**3.7**	**100.0%**	**15.3**	**1.2**	**15.3**	**83.3%**	**27.5**	**2.1**
Δ	-	**+6.5**	**+26.0**	**+16.5**	**+1.4**	**+35.3**	**+73.0**	**+41.6**	**+4.4**

Results. As shown in Table 1, even strong models like LLaVA-OV-7B exhibit significant hallucination rate increases in multi-image scenarios. For example, when input images grow from 4 to 8, *Detail Misinterpretation hallucinations* become severe—the CHAIR score jumps from **10.2 to 50.6**, indicating a sharp decline in grounding accuracy. The sharp drop in the SCover score from **74.0% to 10.3%** also reveals *context omission issues* within multi-image settings. Analysis of failure cases (Fig. 1) reveals that *Context Conflation* commonly occurs, critically degrading model performance.

4 CcDPO: Context-to-Cue Direct Preference Optimization

As discussed in Sec. 3, three fundamental hallucinations (context omission, context conflation, and detail misinterpretation) manifest as performance degradations in MLLMs' multi-image understanding. To address these challenges, we propose **C**ontext-to-**C**ue **D**irect **P**reference **O**ptimization **(CcDPO)**, a hierarchical preference alignment framework that refines MLLMs at two levels (as shown in Fig. 3):

- **Context-Level Optimization:** By contrasting complete and disrupted multi-image captions using language-based preference optimization, we enhance MLLMs' contextual understanding by ensuring comprehensive integration of all relevant visual information across image sequences.

- **Needle-Level Optimization:** A hybrid preference optimization method integrates two complementary objectives: (1) Contrasts captions that either align with or mismatch visually prompted regions using language-based preference optimization, and (2) Discriminates between images semantically matching or contradicting given captions using vision contrastive preference optimization. This dual approach trains the model to make preference judgments grounded in fine-grained visual details.

To support such hierarchical DPO, we construct **MultiScope-42k**, a large-scale preference dataset with automatically generated positive and perturbed response pairs at both levels. See Appendix A.

4.1 Context-Level DPO with Language-Based Preference Optimization

We propose a low-cost DPO mechanism for MLLMs that leverages multi-image captioning as a proxy task to enforce coherent multi-image understanding, thereby addressing context hallucinations (omission, conflation). Specifically, we reformulate the response generation task as a structured, per-image captioning problem. Each image in a sequence is described in an explicit format, generating the preferred response $\boldsymbol{y}_w$ as a coherent sequence of captions that reflect the content of each image:

$$\boldsymbol{y}_w = \{\texttt{For Image 1:}\langle\texttt{caption 1}\rangle\texttt{, For Image 2:}\ldots\texttt{, For Image N:}\langle\texttt{caption N}\rangle\texttt{.}\}$$

This encourages the model to practice selective attention and attribution-aware generation. During training, we construct preference pairs where the **positive sample** is a coherent, full-sequence description $\boldsymbol{y}_w$, and the **negative sample** $\boldsymbol{y}_l$ is obtained from following two perturbation strategies:

- **Sequence Truncation**: simulates context omission by either removing captions entirely from one or more images (complete content omission) or replacing detailed captions with shorter versions (partial content omission), generating rejected responses $\boldsymbol{y}_l^{\text{trunc}}$ and $\boldsymbol{y}_l^{\text{short}}$. Complete omission disrupts sequence continuity, while partial omission results in sparse or incomplete sequence information:

$$\boldsymbol{y}_l^{\text{trunc}} = \{\texttt{For Image 1:}\langle\texttt{caption 1}\rangle\texttt{, For Image 3:}\langle\texttt{caption 3}\rangle\texttt{, }\ldots\}$$
$$\boldsymbol{y}_l^{\text{short}} = \{\texttt{For Image 1:}\langle\text{ short caption 1 }\rangle\texttt{, For Image 2:}\langle\text{ short caption 2 }\rangle\texttt{, }\ldots\}$$

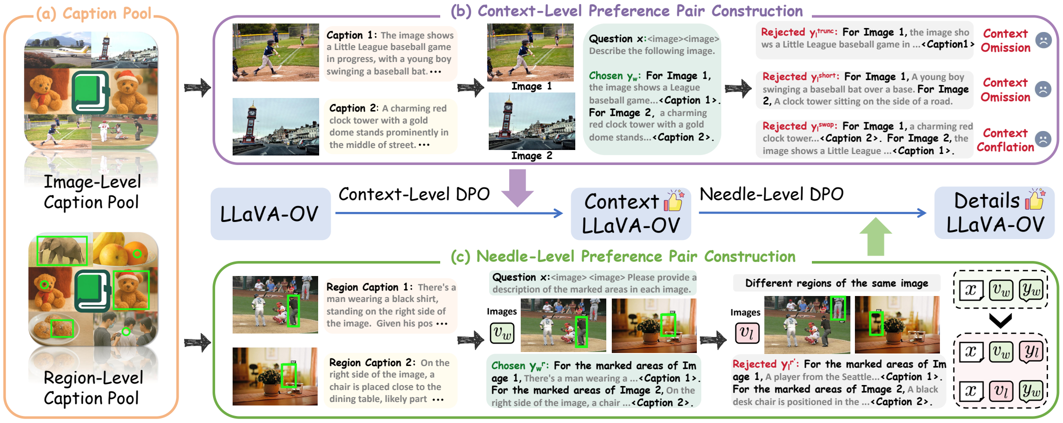

(a) Caption Pool
Image-Level Caption Pool
Region-Level Caption Pool
(b) Context-Level Preference Pair Construction
Caption 1: The image shows a Little League baseball game in progress, with a young boy swinging a baseball bat. ···
Caption 2: A charming red clock tower with a gold dome stands prominently in the middle of street. ···
Image 1
Image 2
Question x:<image><image> Describe the following image.
Chosen y_w: For Image 1, the image shows a League baseball game...<Caption 1>. For Image 2, a charming red clock tower with a gold dome stands...<Caption 2>.
Rejected y_l^trunc: For Image 1, the image sho ws a Little League baseball game in ...<Caption1>
Context Omission
Rejected y_l^short: For Image 1, A young boy swinging a baseball bat over a base. For Image 2, A clock tower sitting on the side of a road.
Context Omission
Rejected y_l^swap: For Image 1, a charming red clock tower...<Caption 2>. For Image 2, the image shows a Little League ...<Caption 1>.
Context Conflation
LLaVA-OV
Context-Level DPO
Context LLaVA-OV
Needle-Level DPO
Details LLaVA-OV
(c) Needle-Level Preference Pair Construction
Region Caption 1: There's a man wearing a black shirt, standing on the right side of the image. Given his pos ···
Region Caption 2: On the right side of the image, a chair is placed close to the dining table, likely part ···
Question x:<image> <image> Please provide a description of the marked areas in each image.
Images
v_w
Chosen y_w^r: For the marked areas of Image 1, There's a man wearing a ...<Caption 1>. For the marked areas of Image 2, On the right side of the image, a chair ...<Caption 2>.
Different regions of the same image
Images
v_l
Rejected y_l^r': For the marked areas of Image 1, A player from the Seattle...<Caption 1>. For the marked areas of Image 2, A black desk chair is positioned in the ...<Caption 2>.
x v_w y_w
x v_w y_l
x v_l y_w

Figure 3: **Overview of CcDPO.** (a) Caption pools are built from LLaVA-23K [61], MDVP [63], and MVC [49] for image- and region-level supervision. (b) Context-Level DPO aligns model outputs with complete, coherent image sequences and penalizes omissions, conflation, and misalignments. (c) Needle-Level DPO incorporates visual prompts to enhance local detail understanding. chosen responses describe marked regions accurately, while rejected are drawn from mismatched regions. Both language-based and vision-contrastive preferences are used to sharpen fine-grained perception.

- **Content Swapping**: simulates context conflation by swapping the captions between different images within the sequence, thereby generating the rejected response $\boldsymbol{y}_l$. e.g.,

$$\boldsymbol{y}_l^{\text{swap}} = \{\texttt{For Image 1:}\langle\texttt{caption 2}\rangle\texttt{, For Image 2:}\langle\texttt{caption 1}\rangle\texttt{, ...}\}$$

Given an instruction $\boldsymbol{x}$ and image sequence $\boldsymbol{v}_w$, we define the context-level DPO objective as:

$$\mathcal{L}_{\text{DPO}_t} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid v_w, x)}{\pi_{\text{ref}}(y_w \mid v_w, x)} - \beta\log\frac{\pi_\theta(y_l \mid v_w, x)}{\pi_{\text{ref}}(y_l \mid v_w, x)}\right), \quad y_l \in \{y_l^{\text{trunc}}, y_l^{\text{short}}, y_l^{\text{swap}}\} \tag{1}$$

where π_θ is the target model and π_{ref} is a frozen reference model. This objective reinforces the preference $(\boldsymbol{x}, \boldsymbol{v_w}, \boldsymbol{y_w}) \succ (\boldsymbol{x}, \boldsymbol{v_w}, \boldsymbol{y_l})$ for globally coherent responses over disrupted ones. We use LLaVA-23K [61] and COCO [62] as our detailed and brief context caption pool, respectively. The visualization cases for three different types of context preferences are presented in Figs. 6, 7, and 8.

4.2 Needle-Level DPO with Hybrid Visual-Language Optimization

Even when global context is preserved, MLLMs often fail to identify or attend to salient visual elements (e.g., missing objects, actions, or attributes). This leads to **detail misinterpretation**, which global response-based supervision alone cannot resolve. To address this, we introduce **needle-level optimization**, a fine-grained hybrid preference mechanism employing visual prompts and image-level perturbations to sharpen the model's focus on often-overlooked local visual cues.

Language-based Preference Optimization (TDPO). This stage leverages region-specific visual prompts (bbox and point) to guide the model's preference judgments. The visual prompt construction involves three key steps: (1) collecting source data with point or bounding box annotations, where points are defined as [x, y] coordinates and bounding boxes as [x1, y1, x2, y2] corner coordinates; (2) converting these annotations into visual prompts (e.g., red points or green rectangles) overlaid onto original images; (3) enhancing region salience by adding textual labels (e.g., "REF") adjacent to visual markers, as illustrated in Fig. 3. These visual prompts explicitly define the target regions for description. To train the model to interpret these specified areas accurately, our DPO method learns from preference pairs that penalize misinterpretations. These pairs are constructed as follows:

1) Chosen Responses ($\boldsymbol{y}_{\boldsymbol{w}}^{r}$): We integrate region-level visual prompts (e.g., bounding boxes, keypoints) into images $\boldsymbol{v}$ to highlight a target region $\boldsymbol{r}$, yielding $\boldsymbol{v}_{\boldsymbol{r}}$. The model is trained to prefer the accurate description $\boldsymbol{y}_{\boldsymbol{w}}^{r}$ of this specific region, directing its attention to critical visual elements:

$$\boldsymbol{y}_w^r = \{\texttt{For the marked area of Image 1: }\langle\texttt{caption } r_1\rangle\texttt{, For the marked area of Image 2: }\langle\texttt{caption } r_2\rangle\texttt{,..., For the marked area of Image N: }\langle\texttt{caption } r_N\rangle\texttt{.}\}$$

2) Rejected Responses ($\boldsymbol{y}_l^{r'}$): Descriptions of regions $\boldsymbol{r'}$ that are non-overlapping with $\boldsymbol{r}$ within the same image serve as rejections. By learning to identify and reject such region-specific inaccuracies, the model's perception of fine-grained image details is more precise. $\boldsymbol{y}_l^{r'}$ is formally defined as:

$$\boldsymbol{y}_l^{r'} = \{\texttt{For the marked area of Image 1: } \langle \texttt{caption } r'_1 \rangle, \texttt{For the marked area of Image 2: } \langle \texttt{caption } r'_2 \rangle, \ldots, \texttt{For the marked area of Image N: } \langle \texttt{caption } r'_N \rangle. \}$$

Given an instruction $\boldsymbol{x}$ and the image sequence $\boldsymbol{v_r}$, these pairs $(\boldsymbol{x}, \boldsymbol{v_r}, \boldsymbol{y}_w^r) \succ (\boldsymbol{x}, \boldsymbol{v_r}, \boldsymbol{y}_l^{r'})$ inform the language-based DPO objective (Eq. 1). We utilize MDVP [63] for the region-level caption pool.

Vision Contrastive Preference Optimization (VDPO). Inspired by [49, 31], this stage further hones the model's visual discrimination. It trains the model by contrasting a single description $\boldsymbol{y_w}$ against two image inputs: $\boldsymbol{v_w}$, which correctly aligns with $\boldsymbol{y_w}$ (often focusing on a specific visual region), and $\boldsymbol{v_l}$, which is misaligned or visually contradicts $\boldsymbol{y_w}$. The objective combines two components:

1) Focusing on Relevant Visuals ($\mathcal{L}_{\text{Focus}}$): This rewards prioritizing details in the correctly aligned image $\boldsymbol{v_w}$ when generating $\boldsymbol{y_w}$, countering MLLMs' tendency to neglect visual content.

$$\mathcal{L}_{\text{Focus}}(v_w, y_w) = -\log \sigma \left(\beta_1 \log \frac{\pi_\theta(y_w \mid v_w, x)}{\pi_{\text{ref}}(y_w \mid v_w, x)} - \beta_1 \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} \right), \tag{2}$$

2) Rejecting Contradictory Visuals ($\mathcal{L}_{\text{Reject}}$): This penalizes assigning high probability to $\boldsymbol{y_w}$ when conditioned on a contradictory image $\boldsymbol{v_l}$.

$$\mathcal{L}_{\text{Reject}}(v_l, y_w) = -\log \sigma \left(\beta_2 \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta_2 \log \frac{\pi_\theta(y_w \mid v_l, x)}{\pi_{\text{ref}}(y_w \mid v_l, x)} \right), \tag{3}$$

The combined vision contrastive DPO loss is $\mathcal{L}_{\text{DPO}_v}(v_w, y_w, v_l) = \mathcal{L}_{\text{Focus}}(v_w, y_w) + \mathcal{L}_{\text{Reject}}(v_l, y_w)$. This objective sharpens the model's ability to distinguish fine-grained visual cues by rewarding focus on relevant details and penalizing attention to misleading content. We use the MVC [49] dataset as a region-level visual counterfactual caption pool. Visualization cases are shown in Figs. 9, 10, and 11.

5 Experiments

5.1 Experimental Settings and Evaluation Benchmarks

Baselines. We apply CcDPO to two different 7B-size MLLMs: Qwen2-VL [33] and LLaVA-OV [32]. Due to differences in base models, preference data, and alignment strategies, direct comparisons with other LLMs are not possible. However, we provide the results for reference: LLaVA-1.5 [64], InternVL2-8B [65], Mantis-Idefics [44], mPLUG-Owl3 [66], Idefics2-8B [20], and Emu2-Chat [21].

Implementation Details. Our model undergoes a three-stage sequential training process to better understand multi-image preferences at both broad (context) and detailed (needle) levels. **Stage 1** focuses on context-level alignment, where we fine-tune Qwen2-VL-7B and LLaVA-OV-7B for one epoch with learning rates of 5×10^{-6} and 5×10^{-5}, respectively, using Eq. 1. **Stage 2** applies needle-level language-based DPO using Eq. 1 to improve sensitivity to fine-grained visual cues with the same learning rate of 5×10^{-5}. We conduct Stage 1 and Stage 2 by using LoRA adaptation [67] with rank $r = 128$ for efficiency. **Stage 3** performs vision contrastive DPO via full-parameter tuning for one epoch with a learning rate of 1×10^{-6} using Eq. 2, strengthening the model's ability to distinguish preferred visual content. Following the setup in [27], we set the temperature parameter $\beta = \beta_1 = \beta_2 = 0.1$ and the negative log-likelihood (NLL) loss coefficient $\gamma = 0.1$. All training is conducted on eight GPUs, each equipped with 90GB of memory.

Evaluation Benchmarks. We employ seven multi-image benchmarks—MUIRBench [42], MIRB [68], BLINK [69], Mantis-Eval [44], NLVR2 [70], Q-Bench2 [71], and MIBench [72]—to holistically evaluate multi-image reasoning across four key dimensions: co-reference alignment, fine-grained comparison, contextual reasoning, and temporal understanding. Complementing these, eight representative single-image benchmarks assess specific multimodal capabilities: (1) Academic/Scientific Reasoning: MMMU [73], MMStar [74], ScienceQA [75], (2) Diagram Understanding: AI2D [76], (3) Robustness against hallucinations: POPE [77], HallBench [38], (4) General Multimodal Abilities: MMBench [78], (5) Text Recognition: OCRBench [79]. This comprehensive evaluation suite demonstrates our method's strengths in both holistic understanding and fine-grained visual grounding across single-image and multi-image general tasks.

Table 2: **Main results on general multi-image benchmarks.** We compare our proposed method, CcDPO, with existing multi-image DPO approaches across seven multi-image benchmarks. Our method consistently enhances the performance of both LLaVA-OV and Qwen2-VL.

Models	**Parameter**	**MuirBench**	**MIRB**	**BLINK**	**Mantis**	**NLVR2**	**MIBench**	**Q-Bench2**	**Average**
GPT-4o [80]	-	62.3	53.0	60.1	62.7	88.8	71.8	74.5	67.6
LLaVA-v1.5 [64]	7B	19.9	28.4	37.1	41.9	52.1	40.9	53.9	39.2
Idefics2 [20]	8B	26.1	33.0	45.2	48.9	86.9	29.7	57.0	46.6
Mantis-Idefics2 [44]	8B	44.5	41.8	49.1	57.1	89.7	44.3	75.3	57.4
mPLUG-Owl3 [66]	8B	39.6	-	50.3	63.1	90.8	54.5	-	59.6
Emu2-Chat [21]	37B	33.6	27.2	36.2	37.8	58.2	39.7	65.3	42.6
InternVL2-8B [65]	8B	48.7	50.0	50.6	60.3	85.56	52.9	-	58.0
LLaVA-OV [32]	7B	42.5	47.3	51.1	60.4	89.4	73.6	73.8	62.5
+ SFT	7B	45.4	48.9	53.4	64.9	89.0	71.9	75.7	64.1
+ MIA-DPO [27]	7B	41.4	48.0	53.7	60.3	88.2	67.8	74.0	61.9
+ CcDPO (Ours)	7B	**48.6**	**51.4**	**55.9**	**69.6**	**91.2**	**75.2**	**77.6**	**67.1**
Δ	-	**+6.1**	**+4.1**	**+4.8**	**+9.2**	**+1.8**	**+1.6**	**+3.8**	**+4.6**
Qwen2-VL [33]	7B	40.5	59.5	53.4	65.9	84.8	68.9	74.5	63.9
+ SFT	7B	43.1	59.8	54.7	64.9	85.2	69.4	74.1	64.5
+ MIA-DPO [27]	7B	40.1	**61.4**	54.5	69.1	84.5	66.7	75.6	64.5
+ CcDPO (Ours)	7B	**44.8**	60.7	**56.5**	**69.1**	**86.4**	**71.9**	**77.0**	**66.6**
Δ	-	**+4.3**	**+1.2**	**+3.1**	**+3.2**	**+1.6**	**+3.0**	**+2.5**	**+2.7**

Table 3: **Main results on single-image benchmarks.** We compare our CcDPO with existing DPO-based approaches across seven single-image benchmarks. Our CcDPO not only improves performance in multi-image settings but also preserves strong capabilities on single-image tasks.

Models	**Parameter**	**MMStar**	**SQA**	**MMMU**	**POPE**	**HallBench**	**MMB**	**OCR**	**AI2D**	**Avg.**
LLaVA-v1.6 [81]	7B	37.6	87.5	35.8	70.3	51.6	69.8	53.7	67.0	59.1
Qwen-VL-Chat [82]	7B	34.5	68.8	35.9	74.9	39.2	61.8	48.8	63.0	53.3
Idefics2 [20]	8B	49.5	88.7	43.0	86.2	-	75.7	-	72.3	69.2
OpenFlamingo [83]	9B	36.9	44.8	-	52.6	38.4	32.4	14.9	31.7	35.9
InstructBLIP [84]	13B	32.7	54.1	-	86.1	45.3	38.3	27.6	40.6	46.3
Emu2-Chat [21]	37B	40.7	68.2	36.3	88.0	-	63.4	43.6	49.7	55.7
LLaVA-OV [32]	7B	58.7	92.1	**47.7**	86.1	52.9	81.8	47.3	81.6	68.5
+ SFT	7B	57.8	91.5	47.1	**88.4**	57.2	81.5	50.2	81.6	69.4
+ MIA-DPO [27]	7B	57.4	92.4	45.1	87.9	55.4	80.9	**52.1**	81.5	69.0
+ CcDPO (Ours)	7B	**59.5**	**92.6**	45.7	86.6	**58.4**	**81.9**	51.0	**82.1**	**69.7**
Qwen2-VL [33]	7B	57.8	**84.1**	50.6	85.9	66.9	81.2	**85.6**	78.9	73.8
+ SFT	7B	55.0	82.7	50.0	87.7	66.7	81.0	84.8	78.5	73.3
+ MIA-DPO [27]	7B	58.2	84.0	48.6	**88.4**	62.7	80.8	85.1	78.9	73.3
+ CcDPO (Ours)	7B	**58.7**	82.8	**50.7**	87.1	**68.8**	**81.6**	83.5	**79.7**	**74.1**

5.2 Main Results

Results on General Multi-Image Tasks. As shown in Table 2, we evaluate CcDPO across diverse multi-image benchmarks that span a wide range of reasoning skills. CcDPO consistently outperforms both the SFT baseline and other DPO-based methods on all datasets, with a notable gain of +4.8 points on BLINK, which focuses on multi-view and spatial reasoning. On the large-scale MuirBench dataset—where each sample contains an average of 4.3 images and up to 9 images—CcDPO achieves the largest improvement of +6.1 points, demonstrating its strength in modeling complex multi-image dependencies such as fine-grained perception, sequential cues, and holistic context. In contrast, MIA-DPO underperforms on MuirBench, highlighting its global context limitations. Consistent gains across LLaVA-OV and Qwen2-VL variants validate our approach's generality and effectiveness.

Results on General Single-Image Tasks. While previous works [27, 44] indicate that multi-image training can degrade single-image understanding, our CcDPO, in contrast, generally yields performance gains on most single-image datasets as shown in Table 3, averaging +1.2 points for LLaVA-OV and +0.3 for Qwen2-VL. Visually-driven tasks like HallBench exhibit the largest improvements, up to +5.5 points under CcDPO. Conversely, for tasks with relatively low reliance on visual information, exemplified by ScienceQA [75], our method showed no notable gains, and performance slightly declined. These results highlight CcDPO's robustness: it not only excels in multi-image scenarios but also preserves, and often enhances, single-image capabilities. We attribute this success to our

Table 4: **Performance on the image retrieval task from needle-in-a-haystack MM-NIAH [85].** We compare our proposed method, CcDPO, with DPO-based baselines across 1K–24K contexts, where the number of images ranges from a few to over a hundred. CcDPO consistently outperforms prior methods, demonstrating its strength in capturing fine-grained details in ultra-long image sequences.

Models	Parameter	1K	2K	4K	8K	12K	16K	24K	Average
LLaVA-OV-7B [32]	7B	89.2	88.1	82.3	71.2	65.0	60.9	45.0	71.7
+ SFT	7B	92.0	93.9	88.2	80.7	74.8	69.4	49.9	78.4
+ MIA-DPO [27]	7B	93.9	94.6	90.5	85.1	75.5	68.8	59.4	81.1
+ CcDPO (Ours)	7B	**95.3**	**96.9**	**91.4**	**89.6**	**78.8**	**74.5**	**64.7**	**84.5**
Δ	-	**+6.3**	**+8.8**	**+9.1**	**+18.4**	**+13.8**	**+13.6**	**+19.7**	**+12.8**

Table 5: **Ablation study of two-level CcDPO on MIBench and MIRB tasks** requiring perception, comparison, and reasoning across multiple images. Detailed task descriptions are in the Appendix B.

	MIBench Benchmark								**MIRB Benchmark**			
Models	**GC**	**SD**	**TR**	**LR**	**FVR**	**TRI**	**VTK**	**TVK**	**Know.**	**Reas.**	**Perc.**	**M-Hop**
LLaVA-OV-7B [32]	87.7	85.9	72.6	74.5	96.5	77.5	42.7	67.1	70.0	44.0	50.0	12.0
+ Context-Level	87.4	88.6	69.5	76.0	97.9	76.7	43.4	68.8	75.0	44.0	52.0	15.0
⊕ Needle-Level	88.8	90.4	70.3	76.0	98.2	77.6	52.6	69.8	72.0	48.0	55.0	18.0
Δ	**+1.1**	**+4.5**	**-2.3**	**+1.5**	**+1.7**	**+0.1**	**+9.9**	**+2.7**	**+2.0**	**+4.0**	**+5.0**	**+6.0**
+ SFT	87.9	88.2	71.0	75.5	87.8	77.4	38.5	68.3	75.0	43.0	53.0	9.0
+ MIA-DPO [27]	85.8	87.2	63.4	67.9	94.9	67.6	42.0	59.6	73.0	50.0	44.0	11.0

preference data design, which employs structured, per-image descriptions, thereby fostering precise understanding of individual images even within multi-image contexts.

5.3 Ablation Studies

Comparison with SFT Trained on MultiScope-42K. As shown in Table 2, Table 3, Table 4, CcDPO outperforms SFT across all benchmarks, achieving +3.0 on multi-image tasks, +0.6 on single-image tests, and a significant +6.1 gain on MM-NIAH. By integrating negative samples into DPO, CcDPO enhances discrimination between accurate and hallucinated outputs, improving fine-grained detail recognition and long-range dependency modeling while maintaining single-image performance. This demonstrates both robust generalization and superior contextual understanding, with negative sample integration proving essential to its performance advantages.

Superior Context Scaling for Fine-Grained Detail Capture. As shown in Table 4, CcDPO significantly outperforms prior DPO-based methods and SFT baselines on the challenging MM-NIAH [85] needle-in-a-haystack image retrieval task. Remarkably, CcDPO achieves consistent improvements across all tested context lengths (1K–24K), with its largest gain (+19.7 points) occurring at the maximum 24K context length compared to the baseline model. This demonstrates CcDPO's exceptional ability to: (1) Scale effectively to ultra-long image sequences, and (2) Capture fine-grained visual details critical for discriminating subtle differences in large image collections. The +12.8-point average improvement underscores its robustness in handling large-scale multi-image contexts. The hierarchical preference optimization in CcDPO enables precise, context-aware understanding—an essential capability for processing extensive visual information.

Effectiveness of Context-Level Optimization. The Context-Level DPO enhances alignment between language responses and the holistic visual context across image sequences. As shown in Table 6, this yields significant improvements on sub-tasks requiring global reasoning, including: Diagram Understanding (+5.1), Image-Text Matching (+9.8), Similarity Matching (+16.4). These gains reflect the module's ability to: (1) Capture semantic relationships across images, and (2) Maintain coherent multi-image descriptions through consistent attribute attribution. Notably, Scene Understanding and Retrieval tasks also benefit from improved global alignment, confirming that such optimization effectively reduces context omission and conflation errors in complex visual sequences.

Effectiveness of Needle-Level Optimization. The Needle-Level DPO improves the model's ability to capture fine-grained visual cues by contrasting localized content. As shown in Table 5 and Table 6, this is especially effective for tasks requiring detailed comparisons across images. In MIBench, the VTK task—where the model must link information across image cells—shows a large gain of +9.9, demonstrating that our visual preference signals help focus on factual visual details. Similarly,

Table 6: **Ablation study of CcDPO on the MuirBench dataset** across all sub-datasets, demonstrating significant performance gains on most subsets. The symbol ⊕ stands for method superposition.

Models	**Overall**	Action	Similarity	Cartoon	Counting	Diagram	Difference	Geographic	I-T Match	Ordering	Scene	Grounding	Retrieval
LLaVA-OV-7B [32]	42.5	35.9	33.1	35.8	24.7	55.0	30.0	46.0	46.9	20.3	72.0	29.7	47.6
+ Context-Level	44.8	36.6	30.3	37.1	35.5	57.0	37.9	41.0	48.3	15.6	65.1	28.3	48.3
⊕ Needle-Level-TDPO	47.8	42.3	48.0	34.6	38.0	58.3	40.3	41.0	55.4	14.1	68.9	30.1	45.9
⊕ Needle-Level-VDPO	48.6	44.5	49.5	33.3	39.3	60.1	39.1	41.0	56.7	13.0	70.4	28.6	46.6
Δ	**+6.1**	**+8.6**	**+16.4**	**-2.5**	**+14.6**	**+5.1**	**+9.1**	**-5.0**	**+9.8**	**-7.3**	**-1.6**	**-1.1**	**-1.0**
Qwen2-VL-7B [33]	40.5	40.8	46.4	41.0	39.7	41.9	34.4	25.0	54.3	9.3	65.0	28.5	20.5
+ Context-Level	42.1	41.0	45.0	42.3	39.8	46.2	37.9	22.0	56.9	14.1	65.1	27.4	21.0
⊕ Needle-Level-TDPO	42.3	39.6	43.9	41.0	39.7	46.5	39.1	22.0	56.3	15.7	66.7	26.2	22.9
⊕ Needle-Level-VDPO	44.8	43.9	46.9	38.5	39.9	53.8	37.9	20.0	61.4	18.8	69.9	23.8	22.9
Δ	**+4.3**	**+3.1**	**+0.5**	**-2.5**	**+0.2**	**+11.9**	**+3.5**	**-5.0**	**+7.1**	**+9.5**	**+4.9**	**-4.7**	**+2.4**

Table 7: Ablation on component and training order.

Configuration	MuirBench	Mantis	BLINK	Q-Bench2
Baseline	42.5	60.4	51.1	73.8
Context-Level Only	44.4	69.0	55.3	76.4
Needle-Level Only	46.3	65.0	54.3	75.5
Needle → Context	46.5	66.3	54.1	**77.8**
Context → Needle	**48.6**	**69.6**	**55.9**	77.6

Table 8: Ablation on visual prompt types.

Bbox	Seg.	Point	Circle	BLINK	Q-Bench2
	✔		✔	55.1	76.0
		✔	✔	56.2	76.6
	✔	✔		**56.8**	76.3
✔			✔	54.5	**77.2**
✔		✔		55.9	76.8

in MuirBench, we observe strong gains on Action Understanding (+8.6), Counting (+14.6), and Difference Spotting (+9.1), all of which depend on localized perception. These results indicate that Needle-Level DPO significantly boosts the model's perceptual grounding and resistance to detail-level hallucinations, complementing the context-level DPO for better multi-image understanding.

Synergistic Two-Level Optimization in CcDPO. As evidenced in Table 7, performing context-level DPO prior to needle-level DPO consistently yields optimal benchmark performance, underscoring a hierarchical interdependence between the two stages. This workflow mirrors human general reasoning for multi-image problems: one first understands the global context, then zooms in to focus on specific details as needed for the task. Notably, inverting the training sequence or isolating either stage results in suboptimal outcomes, further validating the necessity of our two-stage sequential optimization.

Sensitivity Analysis of Visual Prompt. As shown in Table 8, our ablation studies reveal minimal performance variation (±1 point) across visual prompt types (bounding boxes, circles, segmentation masks) during the Needle-Level Language-based Preference Optimization stage, underscoring our method's robustness. Notably, we observe benchmark-specific patterns: fine-grained prompts (e.g., segmentation masks) excel in spatial reasoning tasks BLINK by capturing precise local details, while coarse-grained prompts (e.g., bounding boxes) perform better in image-level comparisons Q-Bench2 through efficient spatial guidance. Consequently, we adopt bounding box + keypoint as the final format, balancing accuracy and annotation practicality optimally.

Limitations. While CcDPO is primarily designed for general multi-image reasoning, it does not explicitly model temporal dependencies, which may limit its performance on video-like inputs. However, our framework is readily extensible to such video data by incorporating video clip descriptions (as demonstrated in Table 13). Similarly, although OCR supervision is limited in our current dataset, CcDPO can be naturally enhanced with targeted text-centric preference data in future work.

6 Conclusion

This work introduces **CcDPO**, a two-level preference optimization method for enhancing multi-image understanding in MLLMs. By decoupling learning into context-level and needle-level stages, CcDPO addresses key hallucination issues including context omission, conflation, and detail misinterpretation. The context-level module promotes holistic sequence comprehension via structured caption preferences, while the needle-level module strengthens fine-grained perception through visual prompts and contrastive supervision. To support optimization, we construct **MultiScope-42k**, a large-scale dataset with automatically generated multi-level preference pairs. Experiments across seven multi-image benchmarks show that CcDPO achieves consistent improvements over SFT and prior DPO variants, confirming its effectiveness in aligning MLLMs with both global and local visual content.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. 62576299, No. 625B2156, No. U21B2037, No. U22B2051, No. U23A20383, No. U21A20472, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305, and No. 62272401), the Natural Science Foundation of Fujian Province of China (No. 2021J06003, No. 2022J06001), and the Fundamental Research Funds for the Central Universities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] Qiong Wu, Wenhao Lin, Yiyi Zhou, Weihao Ye, Zhanpeng Zen, Xiaoshuai Sun, and Rongrong Ji. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. *arXiv preprint arXiv:2411.19628*, 2024.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[7] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[8] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36:70115–70140, 2023.

[9] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478, 2023.

[10] Neil Cohn, Ryan Taylor, and Kaitlin Pederson. A picture is worth more words over time: Multimodality and narrative structure across eight decades of american superhero comics. *Multimodal Communication*, 6(1):19–37, 2017.

[11] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[14] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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*, pages 24185–24198, 2024.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[17] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[18] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[19] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[20] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

[21] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[22] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[23] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

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

[25] Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models. *arXiv preprint arXiv:2504.13122*, 2025.

[26] Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 37:26171–26194, 2024.

[27] Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. 2025.

[28] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. *arXiv preprint arXiv:2503.13792*, 2025.

[29] Guanghao Zhang, Tao Zhong, Yan Xia, Zhelun Yu, Haoyuan Li, Wanggui He, Fangxun Shu, Mushui Liu, Dong She, Yi Wang, et al. Cmmcot: Enhancing complex multi-image comprehension via multi-modal chain-of-thought and memory augmentation. *arXiv preprint arXiv:2503.05255*, 2025.

[30] Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. *arXiv preprint arXiv:2406.18139*, 2024.

[31] Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv preprint arXiv:2501.16629*, 2025.

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[33] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[34] Xudong Li, Runze Hu, Jingyuan Zheng, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Ke Li, Yunhang Shen, Yutao Liu, Pingyang Dai, et al. Integrating global context contrast and local sensitivity for blind image quality assessment. In *Forty-first International Conference on Machine Learning*, 2024.

[35] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024.

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

[37] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2023.

[38] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, 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*, pages 14375–14385, 2024.

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

[40] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

[41] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

[42] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024.

[43] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2024.

[44] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

[45] Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. *arXiv preprint arXiv:2311.18021*, 2023.

[46] Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*, 2024.

[47] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[49] Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *arXiv preprint arXiv:2502.13928*, 2025.

[50] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[51] Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

[52] Songtao Jiang, Yan Zhang, Ruizhe Chen, Yeying Jin, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*, 2024.

[53] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

[54] Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. *arXiv preprint arXiv:2410.14148*, 2024.

[55] Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Yancheng He, Shilong Li, and Bo Zheng. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*, 2024.

[56] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[57] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[58] Xudong Li, Zihao Huang, Yan Zhang, Yunhang Shen, Ke Li, Xiawu Zheng, Liujuan Cao, and Rongrong Ji. Few-shot image quality assessment via adaptation of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10442–10452, 2025.

[59] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023.

[60] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024.

[61] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[62] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[63] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.

[64] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[65] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[66] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024.

[67] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[68] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*, 2024.

[69] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[70] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[71] Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. 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*, 2024.

[72] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*, 2024.

[73] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[74] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[75] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[76] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

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

[78] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[79] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[80] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[81] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[82] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

[83] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[84] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[85] Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024.

[86] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. *arXiv preprint arXiv:2402.13254*, 2024.

[87] Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*, 2024.

Technical Appendices

In this appendix, we provide additional materials to support a more comprehensive understanding of our proposed method and dataset. **In Sec. A**, we detail the low-cost construction pipeline of **MultiScope-42k** and conduct comparative data analysis with MIA-DPO, including token length distributions, word cloud statistics, and supervision source breakdown. We also clarify the image source overlap between training data and benchmarks to ensure fair evaluation. **In Sec. B**, we summarize all benchmarks used in evaluation, including seven multi-image and eight single-image benchmarks. **In Sec. C**, we provide additional experimental results, including ablation studies on training data volume and supervision granularity to assess their impact on model performance. **In Sec. D**, we present qualitative observations and visualizations of preference pairs.

A MultiScope-42k: A Context-to-Cue Captioning DPO Dataset

A.1 Low-cost Question-Answer Pair Construction

Constructing high-quality instruction-response preference pairs for multi-image learning traditionally requires extensive manual annotation, especially when capturing subtle context dynamics or region-level semantics. To address this bottleneck, we design a low-cost, fully automated pipeline for question-answer pair construction, enabling efficient and scalable data generation with broad coverage and controlled distributional properties.

Automated Caption Pool Sampling. We first leverage existing vision-language datasets—LLaVA-23K [61], MDVP [63], and MVC [49]—to construct a diverse caption pool, containing both image-level and region-level descriptions. By decoupling question construction from caption generation, we are able to sample visual contexts and their aligned captions independently, facilitating large-scale composition of input-output examples.

Structured QA Formatting. Given a sampled image or image sequence, we construct templated instructions (e.g., "Describe the following images" or "Please describe the marked area in each image") to form queries. For the corresponding answers, we use structured formats that encourage compositional reasoning and grounding, such as:

```
[ For Image 1:  <caption 1>, For Image 2:  <caption 2>, ...]
```

```
[ For the marked area of Image X: <caption X>, ...]
```

This approach allows flexible variation in image number, visual scope, and response granularity—supporting both context- and region-level supervision.

Controlled Perturbation for Preference Learning. To generate preference pairs without additional labeling, we apply lightweight perturbation strategies to the answer side only. Specifically:

- (1) Truncation and swapping: simulate omissions and misalignments in context-level answers.
- (2) Region mismatches: in needle-level samples simulate detail hallucination.

These perturbations require **no human involvement** yet introduce controlled errors mirroring real-world MLLM failure modes, enabling scalable preference pair generation with low cost.

Efficient Coverage of Diverse Visual Distributions. With stratified control over image domains, scene compositions, and region attributes via automated sampling, we construct MultiScope-42k: a large-scale corpus covering diverse multi-image tasks. Its diversity in visual layout and semantic granularity delivers robust preference supervision across image types and reasoning levels.

Overall, our pair construction strategy eliminates the need for dense manual annotation while producing rich and challenging preference data at scale—offering a practical solution for instruction tuning in multi-image multimodal models.

A.2 Data Analysis

To better understand the characteristics of our preference supervision, we conduct a comparative analysis of MultiScope-42k and MIA-DPO from both lexical and structural perspectives.

Dataset Composition Overview. Summary statistics of the proposed MultiScope-42k are presented in Table 9, which is divided into three core subsets by supervision granularity and image source, totaling 41.8k preference pairs (27.3k context-level, 10.8k needle-level TDPO, and 3.7k needle-level

Table 9: Summary statistics of MultiScope-42k by supervision level and image source, including total pairs, number of images per instance, and average token lengths of chosen and rejected responses.

Level	Image Source	Total	Images Range	Avg. Chosen Len.	Avg. Rejected Len.
Context-Level	COCO-2014	27.3k	[2, 5]	285.24	165.45
Needle-Level-TDPO	COCO-2017	10.8k	[2, 4]	173.35	173.97
Needle-Level-VDPO	Flickr30k	3.7k	[2, 4]	66.09	65.98

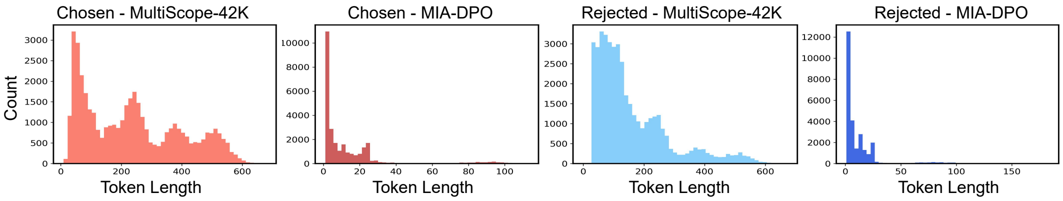


Figure 4: Token length distributions of chosen and rejected responses in our MultiScope-42k and MIA-DPO [27]. MultiScope-42k exhibits significantly longer and more diverse answers, while MIA-DPO responses remain short and concentrated, indicating a simpler response pattern.

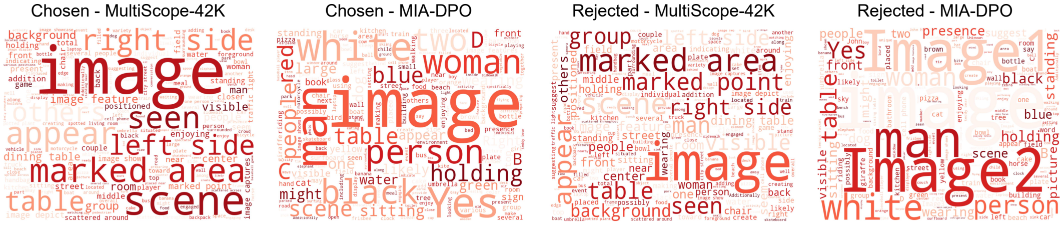


Figure 5: The dataset word cloud comparison between our MultiScope-42k and MIA-DPO [27].

VDPO) to cover supervision needs of varying granularities in multi-image scenarios. The 27.3k context-level pairs incorporate approximately 7k multi-image QA data from Mantis [44] to enhance the model's instruction-following capability. The negative samples are generated by the model based on responses to images with Gaussian noise added, inducing the model to produce answers that are unfaithful to the original images. Regarding the number of images per instance, all subsets contain 2 to 5 images (up to 5 for context-level, 4 for needle-level), aligning with typical multi-image learning settings and supporting models in modeling multi-image sequence contexts.

Token Length Distributions. Fig. 4 presents the token length histograms of both *chosen* and *rejected* responses. MultiScope-42k responses are significantly longer on average and display a wider spread, with many responses exceeding 400 tokens. This reflects the dataset's multi-stage, image-wise captioning format and compositional design. In contrast, MIA-DPO responses are short and concentrated, with the majority under 30 tokens. This further suggests that MultiScope-42k provides richer and more diverse supervision signals, especially for multi-image reasoning.

Linguistic Focus via Word Cloud. We visualize the answer sets of both datasets using word clouds. As shown in Fig. 5, MultiScope-42k responses prominently feature structured and spatially grounded expressions such as *"image," "marked area," "left," "right," "foreground," "scene"*, indicating a strong alignment with multi-image, region-specific prompts. In contrast, MIA-DPO emphasizes atomic visual concepts (e.g., *"man," "table," "dog," "red"*), which are well-suited for single-image tasks but lack explicit inter-image reference or structural composition.

A.3 Dataset Independence Statement

To ensure fair and unbiased evaluation, we verify that the benchmarks used for testing do not overlap with the data sources involved in model training. We employ three types of annotated data for two-level DPO training. (1) Image-level captions are sourced from **LLaVA-23K** [61], which is derived from the COCO-2014 dataset. (2) Region-level captions come from the **MDVP** dataset [63], based on COCO-2017. (3) Visual contrastive preference pairs are taken from **MVC** [49], whose image sources include **CounterCurate** [86] and **FineCops-Ref** [87].

We conducted a thorough review of all evaluation benchmarks for potential overlap with these training sources (COCO-2014, COCO-2017, Flickr30k). While most multi-image benchmarks appear to be independently constructed, we conservatively flag the following for partial or uncertain overlap:

MIRB [68] explicitly uses COCO images (e.g., for counting), and includes content from ImageNet-R, OpenFoodFact, Bitton et al., and arXiv. **MIBench** [72] comprises multiple public datasets, including VrR-VG, which inherits COCO images via Visual Genome. **MUIRBench** [42] integrates existing (e.g., SeedBench, IconQA), derived (e.g., NLVR2, MMBench), and newly collected datasets. While its new data is COCO-free, MMBench and IconQA are known to include COCO images.

A summary of benchmark image sources is provided below:

MUIRBench combines new and repurposed datasets; some derived components include COCO. **MIRB** includes COCO, ImageNet-R, and other web-sourced content. **BLINK** uses synthetic and controlled real-world images. **Mantis-Eval** draws from web queries and manual composition, with no COCO usage. **NLVR2** uses Flickr images and is COCO-independent. **Q-Bench2** is based on IQA datasets like KonIQ-10k and BID, unrelated to COCO or Flickr30k. **MIBench** includes VrR-VG, which partially overlaps with COCO.

While a few benchmarks partially overlap with COCO or Flickr30k, evaluation fairness remains uncompromised for two reasons. First, our primary baseline MIA-DPO is also trained on LLaVA-23K (COCO-based), ensuring all methods are compared on equal footing. Second, our DPO training focuses on image-caption preference alignment, not question answering. These points validate the integrity and fairness of our evaluation protocol, despite partial dataset overlaps.

B Benchmark Sources

B.1 More Details on the Construction of Context-AMBER-1K

We design a caption generation task as a proxy to systematically evaluate MLLMs' multi-image understanding, identifying three fundamental hallucination types—*Context Omission*, *Context Conflation*, and *Detail Misinterpretation*—that critically degrade model performance. To enable this evaluation, we construct **Context-AMBER-1K** by systematically concatenating images from the single-image AMBER dataset [37] into sequences of two types: short-context sequences with 4 images and long-context sequences with 8 images. Each input is paired with the prompt: “`Please sequentially describe each of the images shown above. Use the format: For Image *:<description>.`”.

To evaluate **Detail Misinterpretation**, we employ three established metrics: (a) CHAIR [36]; (b) Response-level hallucination rate (Hal); and (c) Cognition-based hallucination (Cog). Each generated caption is matched to its corresponding image using regular expressions to extract per-image descriptions. We then compute hallucination scores for each image separately and average these scores across the entire image sequence. To assess **Context Omission** and **Context Conflation**, we introduce two rule-based penalization strategies:

(1) For context omission, if fewer than N captions are generated (e.g., the response ends prematurely at `{For Image N-1: <caption N-1>}`), we explicitly pad the output to `{For Image N: <caption N-1>}`) to maintain a consistent structure. Missing or incomplete captions are heavily penalized. Additionally, we introduce the (d) Sequence Coverage Rate (SCover), a metric that evaluates caption completeness across image sequences and quantifies the degree of context omission.

(2) For context conflation, we include the keyword “`sequentially`” in the prompt to encourage models to describe images in order. If the model cannot clearly associate each description with its corresponding image (e.g., `For Image 1:<caption1>, For Image 3:<caption2>, For Image 2:<caption3>`), the resulting hallucination scores will be higher for out-of-order descriptions compared to their correct GTs, as the mismatch leads to a larger discrepancy.

B.2 Multi-Image Benchmarks

We employ seven multi-image benchmarks—MUIRBench [42], MIRB [68], BLINK [69], Mantis-Eval [44], NLVR2 [70], Q-Bench2 [71], and MIBench [72]—to holistically evaluate multi-image reasoning across four key dimensions: co-reference alignment, fine-grained comparison, contextual reasoning, and temporal understanding.

MUIRBench [42] is a comprehensive benchmark specifically designed to evaluate the robustness of multimodal large language models (MLLMs) in multi-image understanding scenarios. It comprises 2,600 multiple-choice questions and 11,264 images, averaging 4.3 images per instance. The benchmark covers 12 distinct multi-image understanding tasks—including action understanding, diagram reasoning, geographic comprehension, and visual retrieval—spanning 10 diverse multi-image relation types such as temporal, narrative, and scene-multiview relations. To ensure both comprehensiveness and robustness, MUIRBench adopts a pairwise design: each standard (answerable) question is paired with an unanswerable variant with minimal semantic perturbations. This enables fine-grained assessments of both reasoning capability and abstention behavior.

MIRB [68] is a comprehensive benchmark designed to evaluate vision-language models (VLMs) on four distinct aspects of multi-image understanding: perception, visual world knowledge, reasoning, and multi-hop reasoning. It comprises 925 multi-image questions across these categories, averaging 3.78 images per question, with some tasks requiring up to 42 images for complex reasoning. Unlike prior benchmarks that reuse video frames, MIRB independently sources images from real-world domains, such as code snippets, sightseeing scenes, food ingredient lists, and arXiv papers, ensuring diverse and challenging visual contexts. The benchmark includes a wide array of tasks: image jigsaw reconstruction, object counting, attribute matching **(Perception)**; food label comparison and geographic recognition **(Knowledge)**; visual analogy, code understanding, 3D scene analysis **(Reasoning)**; and synthetic logic chains and citation lookups **(Multi-Hop)**. Each question is formulated to necessitate reasoning across multiple images rather than from a single image.

MIBench [72] is a large-scale benchmark designed to comprehensively evaluate the fine-grained multi-image understanding abilities of multimodal large language models (MLLMs). It categorizes multi-image inputs into three representative scenarios—Multi-Image Instruction (MII), Multimodal Knowledge-Seeking (MKS), and Multimodal In-Context Learning (MIC)—covering a total of 13 distinct tasks and 13,000 annotated samples. In the MII setting, the model must perform perception, comparison, and reasoning over multiple images across five tasks: general comparison **(GC)**, subtle difference **(SD)**, visual referring **(VR)**, temporal reasoning **(TR)**, and logical reasoning **(LR)**. The MKS scenario evaluates the model's ability to extract and align information from interleaved image-text knowledge sources through four tasks: fine-grained visual recognition **(FVR)**, text-rich image VQA **(TRI)**, vision-linked textual knowledge **(VTK)**, and text-linked visual knowledge **(TVK)**. Finally, the MIC setting assesses multimodal in-context learning across four tasks, including close-ended and open-ended VQA, hallucination mitigation, and demo-based task learning.

BLINK [69] tests rapid visual cognition through perceptual similarity, forensic analysis, and spatiotemporal matching. It includes tightly-controlled multi-image tasks such as depth estimation, object matching, and outlier detection, with an emphasis on speed and perceptual accuracy.

Mantis-Eval [44] introduces 217 multi-image tasks curated for conceptual inference, including abstract reasoning over physical quantities such as number, size, and weight. It combines both multiple-choice and open-ended questions, drawing from web-sourced image sets manually organized into logical visual groupings.

NLVR2 [70] (Natural Language Visual Reasoning) assesses a model's ability to verify textual hypotheses against a pair of images. Each sample requires binary classification (True/False) over whether the provided statement is consistent with both images, making it a canonical test for visual entailment and compositional reasoning.

Q-Bench2 [71] is a diagnostic benchmark tailored for evaluating visual quality perception and comparative assessment across image sets. It challenges models to identify subtle visual artifacts, distortions, or improvements between similar images. Our evaluation is based on the Q-Bench2-A1-dev subset, which emphasizes multi-image multiple-choice assessments for perceptual judgment.

B.3 Single-Image Benchmarks

We test the model on eight representative single-image benchmarks assess specific multimodal capabilities: (1) Academic/Scientific Reasoning: MMMU [73], MMStar [74], ScienceQA [75], (2) Diagram Understanding: AI2D [76], (3) Robustness against hallucinations: POPE [77], HallBench [38], (4) General Multimodal Abilities: MMBench [78], (5) Text Recognition: OCRBench [79]. The results

Table 10: **Benchmark Sources.** We have included detailed information for all the multi-image and single-image benchmarks tested in the paper in the table.

Setting	Models	Evaluation Metric	Number	Source
Multi-Image Benchmark	MUIRBench [42]	Multiple Choice	2,600	**MUIRBench**
	MIRB [68]	Multiple Choice	925	**MIRB**
	MIBench [72]	Multiple Choice	13,000	**MIBench**
	BLINK [69]	Multiple Choice	3,807	**BLINK**
	NLVR2 [70]	Multiple Choice	6,967	**NLVR2**
	Q-Bench2 [71]	Multiple Choice	1,000	**Q-Bench2**
	Mantis-Eval [44]	Multiple Choice	217	**Mantis-Eval**
Single-Image Benchmark	MMStar [74]	Multiple Choice	1,500	**MMStar**
	MMMU [73]	Multiple Choice	1,050	**MMMU**
	Sci-QA [75]	Multiple Choice	4,241	**ScienceQA**
	POPE [77]	Yes/No	9,000	**POPE**
	HallBench [38]	Yes/No	951	**HallusionBench**
	MMB [78]	Multiple Choice	1,164	**MMBench**
	OCR [79]	VQA	1,000	**OCRBench**
	AI2D [76]	Multiple Choice	3,090	**AI2D**

on this diverse set of benchmarks demonstrate the effectiveness of the proposed method, particularly in multi-image settings, confirming significant performance improvements.

MMMU [73] (Massive Multimodal Multitask Understanding) includes over 10k university-level questions from 30+ disciplines such as physics, medicine, and art. It requires detailed reasoning over image-text inputs and is designed to evaluate advanced academic-level understanding.

MMStar [74] is a comprehensive diagnostic benchmark covering various sub-tasks such as OCR, VQA, and caption grounding, offering structured and hierarchical annotations across domains like natural science, medicine, and design.

ScienceQA [75] contains over 21k science questions aligned with elementary and middle school curricula, involving images such as diagrams and charts. It tests the model's capability to perform science-related visual reasoning in a multimodal format.

AI2D [76] (Allen Institute Diagram) features manually annotated science diagrams with associated multiple-choice questions. It focuses on assessing the model's understanding of labeled structures and their functional roles within the image.

POPE [77] (Position and Object-level Prompt Evaluation) is designed to test a model's resistance to hallucinations. It uses minimally perturbed prompts to identify failure cases in positional grounding and object identification, highlighting model robustness.

HallBench [38] provides a structured framework to measure hallucination frequency and grounding quality by comparing model outputs with annotated ground truths. It supports fine-grained scoring across categories such as incorrect object mentions or unsupported claims.

MMBench [78] is a general-purpose evaluation benchmark comprising questions across 11 modalities including VQA, captioning, OCR, and commonsense reasoning. It uses GPT-4-based grading to ensure high-fidelity evaluation of answer correctness.

OCRBench [79] specifically targets the model's capability to recognize and reason about text in the visual domain, covering a range of document layouts, fonts, and multilingual content with both exact-match and reasoning-based questions.

C More Experiments

Data Scale Alignment with MIA-DPO. To assess the impact of training size and ensure a fair comparison with MIA-DPO, we conduct an ablation in Tab. 11 using a similar total number of preference pairs. Specifically, we randomly sample 13.6k from our 27.3k Context-Level pairs and combine them with the fixed 14.5k Needle-Level data, resulting in a 28.1k training set—comparable to MIA-DPO's 28.9k. Notably, under this matched training data size, our **CcDPO** still outperforms

Table 11: **Ablation on training data volume.** To match MIA-DPO's training data size, we downsample our Context-Level data to 13.6k while keeping needle-level data fixed. The results reveal a trade-off between modeling global context and capturing fine-grained details: while reduced Context-Level data leads to performance drops on most multi-image tasks, benchmarks like BLINK and Q-Bench2—focused on localized perception—benefit from a higher proportion of needle-level data.

Models	**Data Size**	**MuirBench**	**MIRB**	**BLINK**	**Mantis**	**NLVR2**	**MIBench**	**Q-Bench2**	**Average**
LLaVA-OV [32]	-	42.5	47.3	51.1	60.4	89.4	73.6	73.8	62.5
+ MIA-DPO [27]	28.9K	41.4	48.0	53.7	60.3	88.2	67.8	74.0	61.9
+ CcDPO (Ours)	28.1K	46.7	51.2	**56.5**	69.1	90.7	72.1	**79.3**	66.5
+ CcDPO (Ours)	41.8K	**48.6**	**51.4**	55.9	**69.6**	**91.2**	**75.2**	77.6	**67.1**
Δ	-	**+6.1**	**+4.1**	**+4.8**	**+9.2**	**+1.8**	**+1.6**	**+3.8**	**+4.6**

Table 13: **Performance on VideoMME Benchmark.** CcDPO-Video improves baseline performance across core video tasks, verifying cross-domain generalizability.

Model	**Overall**	**Perception**					**OCR Problems**
		Temporal	**Spatial**	**Attribute**	**Action**	**Object**	
LLaVA-OV [32]	53.7	0.491	0.556	0.685	0.505	0.559	**0.568**
+ CcDPO-Video (Ours)	**54.4**	**0.509**	**0.574**	**0.698**	**0.527**	**0.582**	0.547
Model	**—**	**Reasoning**					**Information Synopsis**
		Counting	**Temporal**	**Spatial**	**Action**	**Object**	
LLaVA-OV [32]	—	0.354	0.384	0.714	**0.523**	**0.518**	**0.675**
+ CcDPO-Video (Ours)	—	**0.377**	**0.401**	**0.732**	0.519	0.511	0.663

MIA-DPO across all benchmarks, demonstrating the effectiveness of our structured, dual-level supervision. In particular, the reduced-context setting yields better performance on BLINK and Q-Bench2, suggesting that a relatively higher proportion of needle-level data may benefit fine-grained perceptual tasks. On the other hand, performance on context-heavy benchmarks drops slightly, likely due to weaker global context modeling. Overall, training with the full 42k preference set (28.1k Context-Level + 14.5k Needle-Level) leads to the best average performance. These results highlight the advantage of high-quality, large-scale supervision, while also revealing a trade-off between contextual alignment and perceptual precision.

Ablation of Training Strategies. We investigate whether our multi-stage training strategy—first training on Context-Level data, followed by Needle-Level supervision—is more effective than a single-stage approach that mixes both types of data from the beginning. As shown in Tab. 12, the multi-stage strategy consistently outperforms the mixed-data alternative across benchmarks. We attribute this improvement to the sequential learning structure. In the first stage, the model learns to capture global context and image-level coherence through structured, sequence-aligned supervision. Once this foundation is established, the second stage refines the model's ability to attend to local, fine-grained visual cues via region-specific preference optimization. In contrast, the one-stage strategy may dilute the model's focus by simultaneously exposing it to competing global and local objectives, making optimization less efficient. These results suggest that decoupling context modeling and fine-grained grounding into separate stages can better guide the model toward hierarchical visual reasoning.

Table 12: **Ablation on training strategy.** Multi-stage training outperforms one-stage mixed training.

Strategy	**MuirBench**	**BLINK**	**Mantis**
LLaVA-OV	42.5	51.1	60.4
One-Stage	46.2	55.1	65.4
Multi-Stage	**48.5**	**55.9**	**69.6**

Framework Extension to Video Data. To explore our framework's potential for sequential visual data beyond multi-images, we extend its core idea of stitching individual images into multi-image sequences to video scenarios, adopting the structured input format: $\langle$video clip$_1\rangle$ $\langle$caption$_1\rangle$ $\langle$video clip$_2\rangle$ $\langle$caption$_2\rangle$. We directly adapt our **Sequence Truncation** and **Content Swapping** strategies for negative sample generation, avoiding costly manual re-labeling or rejection sampling. Following this pipeline, we construct a 7k-scale video preference dataset and perform DPO training on the LLaVA-OV baseline (denoted as **CcDPO-Video**).

As shown in Table 13, CcDPO-Video achieves a **0.7-point overall gain** over LLaVA-OV on VideoMME, with measurable improvements in temporal perception (+0.018), action recognition (+0.022), and object recognition (+0.023). These exploratory results highlight two promising strengths of our approach: (1) Video clip stitching—an extension of our multi-image sequence construction paradigm—is **computationally economical**, obviating the need for manual re-annotation or heavy computational overhead associated with existing methods; (2) Its perturbation-based sampling is **transferable** to sequential visual data: positive samples use semantically ordered video clips, while negatives are generated via scrambling or truncation. This trial thus demonstrates a feasible pathway to extend multi-image DPO frameworks to video understanding tasks.

D More Observations

As illustrated in Fig. 6, 8, 7, 10, and 11, we present additional qualitative examples of the constructed preference pairs used in our CcDPO training. These examples cover a range of perturbation types across both Context-Level and Needle-Level supervision, each designed to target specific failure modes in multi-image reasoning.

Fig. 6 shows a context-level *complete content omission* scenario, where the rejected response omits part of the image sequence. This simulates a context omission error, encouraging the model to favor complete and coherent descriptions across multiple images.

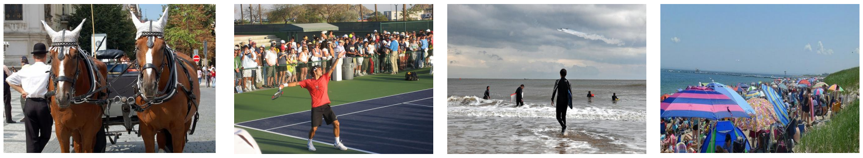
Fig. 7 demonstrates a context-level *partial content omission* scenario, where the rejected response uses overly brief or generic captions. Such examples train the model to prefer rich, fine-grained visual descriptions that capture meaningful details.

Fig. 8 presents a case of *caption swapping*, where descriptions are mismatched with image indices. This context conflation perturbation helps the model learn to correctly associate content with its corresponding image.

Fig. 9 and Fig. 10 move to the Needle-Level and show a case where region-specific (bbox or point) captions are mismatched. This encourages the model to distinguish between accurate and inaccurate local grounding within a given image.

Fig. 11 illustrates the use of contrastive image inputs for the same caption, where one image visually aligns with the text while the other introduces a subtle contradiction. This visual preference pair sharpens the model's ability to resolve ambiguous references through visual evidence alone.

Together, these examples highlight how different types of preference signals are systematically constructed to improve the model's ability to reason over sequences, resolve visual references, and discriminate between subtle semantic mismatches. By explicitly encoding both global and local reasoning objectives into the supervision data, CcDPO promotes more robust and interpretable multi-image alignment.



Question: What is this photo about?<image><image><image><image>

🙂

Chosen: **For image 1:** The image depicts two horses pulling a carriage on a street. Both horses are beautifully decorated, and one of them appears to be wearing chain mail hats. They are tethered to the carriage and working in unison to transport it through the street.\n\nThere are several people in the scene, all around the horses and carriage. Some of the people are walking along the sidewalk, while others may be getting on or off the carriage. A stop sign is located on the ride side of the scene, indicating a controlled intersection. \n

For image 2: The scene shows a man on a tennis court ready to serve the tennis ball. He is holding a tennis racket in his hand, with the ball positioned above him, preparing for the hit. Several spectators surround the court area, attentively watching the man's performance.\n\nOther items can be seen around the court, such as a suitcase located on the right side and a backpack towards the left. Many of the spectators are focused on the tennis player, eagerly anticipating the start of the game. \n

For image 3: The image shows a group of people wearing wet suits at a beach, all carrying surfboards under their arms while in the water. Some surfers are farther out in the water, while one person can be seen walking towards the water with their board. \n\nIn the background, multiple airplanes are flying in the sky, and a small boat is noticeable near the middle of the scene. The bodyboarders seem to be having a great time enjoying their water activities, with the presence of clouds contributing to the beach atmosphere. \n

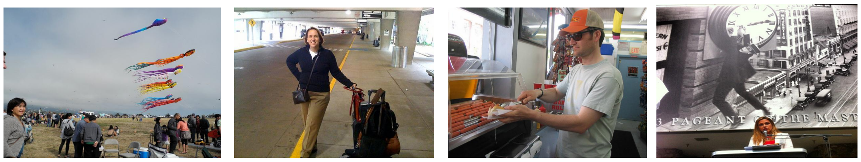
For image 4: The image is of a crowded beach on a clear day with a blue sky. There are numerous people enjoying their time by the ocean, some standing or walking around, while others have set up colorful umbrellas and chairs to relax on the beach. \n\nMany large umbrellas are spread along the shoreline near the water, providing shade for beachgoers, and brightening up the scene with their vibrant colors. A variety of chairs can be found in the area, some near the umbrellas and others scattered across the beach. \n\nA few surfboards and handbags are also visible, indicating that people are likely engaging in water activities and have brought personal belongings for their day at the beach.

😭

Rejected: **For image 1:** The image depicts two horses pulling a carriage on a street. Both horses are beautifully decorated, and one of them appears to be wearing chain mail hats. They are tethered to the carriage and working in unison to transport it through the street.\n\nThere are several people in the scene, all around the horses and carriage. Some of the people are walking along the sidewalk, while others may be getting on or off the carriage. A stop sign is located on the ride side of the scene, indicating a controlled intersection. \n

For image 4: The image is of a crowded beach on a clear day with a blue sky. There are numerous people enjoying their time by the ocean, some standing or walking around, while others have set up colorful umbrellas and chairs to relax on the beach. \n\nMany large umbrellas are spread along the shoreline near the water, providing shade for beachgoers, and brightening up the scene with their vibrant colors. A variety of chairs can be found in the area, some near the umbrellas and others scattered across the beach. \n\nA few surfboards and handbags are also visible, indicating that people are likely engaging in water activities and have brought personal belongings for their day at the beach.

Figure 6: Example of Context-Level preference pair with truncation perturbation (Context Omission).



Question: What is this photo about?<image><image><image><image>

Chosen: **For image 1:** The image features a large group of people gathered at a sandy beach, flying various kites in the cloudy sky. There are more than ten people visible in the scene, dispersed throughout the beach area. The kites come in assorted shapes and sizes, with some resembling squid and octopus wind streamers soaring above the crowd. \n\nThere are also multiple chairs placed in the scene, accommodating those who want to take a break from kite flying or simply enjoy the spectacle. Additionally, a few handbags are lying on the beach, close to their respective owners. \n

For image 2: In the image, a woman is standing outside near a curb with her luggage. She has her hand on her hip, looking confident or perhaps waiting for someone. Her luggage consists of a suitcase next to her and a handbag placed closer to her. \n\nThe scene appears to be near an airport or a drop-off/pick-up area, with several cars parked or driving nearby. A bench is located on the far right of the scene, commonly found at such locations. Additionally, there are multiple other people and various sized handbags and suitcases spotted in the area, indicating that this is a bustling and busy place for travelers. \n

For image 3: The image features a man inside a convenience store or restaurant, in the process of placing a hot dog in a bun. He is surrounded by an assortment of hot dogs in various positions, indicating the convenience store or restaurant has a wide selection of hot dogs for customers to choose from. The man seems focused on preparing his chosen hot dog for consumption. \n

For image 4: A woman is standing on stage, addressing her audience with a microphone. She is wearing a white and red outfit, and there is a large black and white picture in the background. The woman appears to be giving a lecture or making a presentation to her students or a diverse audience.\n\nThere are several people scattered throughout the image, indicating a sizeable gathering listening to the woman's lecture. In the vicinity, various cars are parked outside the venue, some with a bit of distance between them. There is even a truck parked further out near some other vehicles. Additionally, there is a clock visible high up on a wall.",

Rejected: **For image 1:** Crowds of people are at the beach with squid and octopus wind streamers above.

For image 2: A woman standing at an airport next to luggage.

For image 3: The man is at the convenience store selecting a hotdog.

For image 4: The woman is speaking to her audience on stage. ",

Figure 7: Example of Context-Level preference pair with caption shortening perturbation (Context Omission).

Question: What is this photo about?<image><image><image>

🙂 **Chosen:** **For image 1:** The image showcases a bathroom illuminated by a stained glass window, with a white bathtub located next to a stone wall. The room features stone walls and floors, giving the bathroom a rustic feel.\n\nIn the bathroom, a sink is situated along the wall, while a wooden chair is placed in the foreground. Several bottles and cups can also be found near the sink, likely containing bathroom essentials, such as shampoo, soap, and other toiletries. \n

For image 2: The living room is elegantly arranged with old-fashioned, antique furniture and features exquisite details in the woodwork. There is an ornate couch situated on the left side of the room, and several antique style chairs are placed around a dining table. The parlor features an oriental rug, adding an extra layer of charm to the room.\n\nThe living room is also adorned with various decorations such as multiple vases and a potted plant, which accentuate the antique atmosphere.

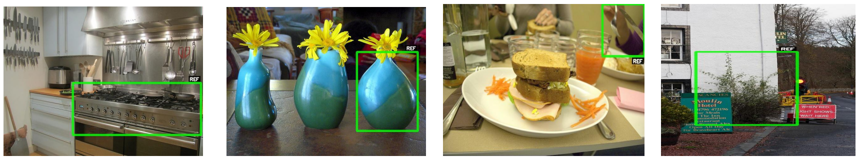
For image 3: In the image, a group of people, including both men and women, is gathered in a park flying kites. Among them, two people are prominently seen flying a red kite in the field, possibly running to keep it airborne. The audience of onlookers is spread out in the park, watching and enjoying the kite-flying activities.\n\nSome chairs are placed around the park for people to sit and relax. Additionally, there are a few traffic lights visible, indicating that the park is likely situated near a road or a busy area.

😭 **Rejected:** **For image 1:** The image showcases a bathroom illuminated by a stained glass window, with a white bathtub located next to a stone wall. The room features stone walls and floors, giving the bathroom a rustic feel.\n\nIn the bathroom, a sink is situated along the wall, while a wooden chair is placed in the foreground. Several bottles and cups can also be found near the sink, likely containing bathroom essentials, such as shampoo, soap, and other toiletries. \n

For image 2: In the image, a group of people, including both men and women, is gathered in a park flying kites. Among them, two people are prominently seen flying a red kite in the field, possibly running to keep it airborne. The audience of onlookers is spread out in the park, watching and enjoying the kite-flying activities.\n\nSome chairs are placed around the park for people to sit and relax. Additionally, there are a few traffic lights visible, indicating that the park is likely situated near a road or a busy area.

For image 3: The living room is elegantly arranged with old-fashioned, antique furniture and features exquisite details in the woodwork. There is an ornate couch situated on the left side of the room, and several antique style chairs are placed around a dining table. The parlor features an oriental rug, adding an extra layer of charm to the room.\n\nThe living room is also adorned with various decorations such as multiple vases and a potted plant, which accentuate the antique atmosphere.

Figure 8: Example of Context-Level preference pair with swapping perturbation (Context Conflation).



Question: <image><image><image><image>
Please provide a detailed description of the marked areas in each image. The marked areas are indicated by a green rectangle with a 'REF' label around it.",

Chosen: **For the marked areas of image 1:** A large commercial-grade stove crawls across a significant section of the kitchen, sparking with arrays of burners. Numerous knobs dot the front, lending a look of complexity and functionality to the appliance. \n

For the marked areas of image 2: On the far right of the image, there's a notably round, possibly teardrop-shaped, vase. This vase seems to be the bulkiest among all, being referred to as the 'fat one'. It is situated in front of some objects which appear to be red. \n

For the marked areas of image 3: An arm wearing a purple sleeve is seen in the top right corner of the image, with a hand gripping a fork poised to cut into some food. \n

For the marked areas of image 4: A tall, lush potted plant can be observed amidst a setting of a white building. This plant, which could be described as a bush or large shrub, is positioned between green and red signs. It appears as a significant green element between these signs, contributing to the natural allure of the scene.

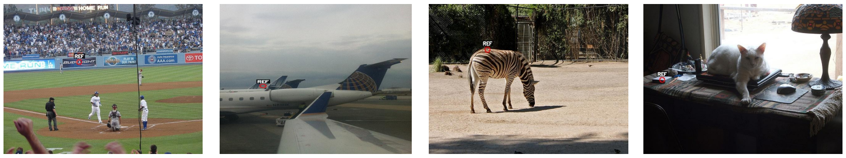
Rejected: **For the marked areas of image 1:** The right part of a stainless-steel oven door is visible in the corner of the image. It almost seems cut-off from the frame and includes a sturdy handle, only partially seen due to positioning. \n

For the marked areas of image 2: At the center of the image, there's a blue vase, distinctively sandwiched between two other vases. Its location in the very center of the image makes it a reference point, much like '12 noon' on a clock. \n

For the marked areas of image 3: A woman in a grey shirt is seen at the top of the image with her left hand gracefully placed on a sandwich. She garners further attention due to the shiny ring on her finger, indicating her fidelity to her husband. \n

For the marked areas of image 4: On the bottom left, there's a noticeably yellowish-green bush, planted in a pot. It's positioned interestingly, partially covering a blue sign and situated next to a sign indicating 'Moulin Hotel'. The position and color of this bush add an element of intrigue and a splash of green to the urban scene.

Figure 9: Example of Needle-Level preference pair with bbox region mismatches perturbation (Detail Misinterpret).



Question:<image><image><image><image>Please provide a detailed description of the marked point in each image. The marked points are indicated by a red hollow circle with a 'REF' label positioned above the circle.

🙂

Chosen: **For the marked point of image 1:** The region shows a segment of a baseball field, focusing on the batter's area by home plate. Visible are players in uniforms engaged in the game, with one in stance to bat and a catcher and umpire in position behind him. The dirt-covered ground, home plate, and the players' equipment, like the batter's helmet, are prominent. \n

For the marked point of image 2: This region shows a commercial airplane tail with distinctive circular branding. The aircraft is stationary on the tarmac, and the tail appears prominently against a cloudy sky. \n

For the marked point of image 3: This is an image of a zebra positioned in profile, grazing on the ground. It is a full-grown adult with characteristic black and white striped fur. The mark is placed near the upper region of the zebra's body, close to the base of its neck. This particular stripe pattern is unique to each individual zebra, similar to a human fingerprint. The immediate area around the mark shows the zebra's mane, which consists of short, erect fur that follows along the neck's curve. \n

For the marked point of image 4: A notepad likely containing handwritten notes or drawings, surrounded by creative or writing utensils.

😭

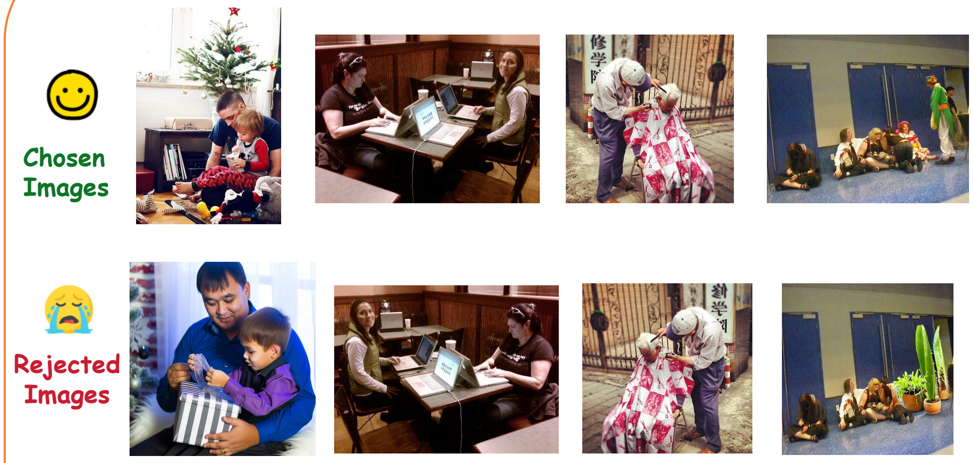
Rejected: **For the marked point of image 1:** This is a display of a large, horizontal electronic advertisement board located at the edge of the stands in a baseball stadium, showing a commercial message or sponsorship. The board appears to be illuminated and is part of the stadium's infrastructure, meant for the audience's viewing. \n

For the marked point of image 2: Focus is on the wingtip of an airplane, showing the curved end and navigation lights. The wing is against a backdrop of the sky ranging from pale to darker shades, possibly indicating cloudy weather.

For the marked point of image 3: This is an image of a zebra positioned in profile, grazing on the ground. It is a full-grown adult with characteristic black and white striped fur. The mark is placed near the upper region of the zebra's body, close to the base of its neck. This particular stripe pattern is unique to each individual zebra, similar to a human fingerprint. The immediate area around the mark shows the zebra's mane, which consists of short, erect fur that follows along the neck's curve. \n

For the marked point of image 4: A small, possibly decorative bowl that may be utilized for holding various small items or as a decor piece itself.

Figure 10: Example of Needle-Level preference pair with point region mismatches perturbation (Detail Misinterpret).



Question: <image><image><image><image>\n
For image 1, what can you tell me about the colors of the child's clothing while they're opening the gift, and can you describe the scene around them?**\n For image 2,** how is the thinner woman positioned in relation to the person in the black shirt in the image?**\n For image 3,** what can you tell me about the position of the elderly Asian man with white hair in relation to the Chinese characters in the image? Is he on the left or the right?**\n For image 4,** how many people are there in the image, and how many costumes can you see?

Chosen: For image 1, The child is dressed in red and gray while opening a gray and white striped gift, sitting on a man's lap who is wearing a blue shirt, and together they are positioned in front of a window and a small Christmas tree.\n

For image 2, The thinner woman is positioned to the right of the person in the black shirt.\n

For image 3, The elderly Asian man with white hair is to the right of the Chinese characters.\n

For image 4, There are 5 people and 6 costumes.

Figure 11: Example of Needle-Level preference pair with image contrastive perturbation (Detail Misinterpret).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: In the introduction, I clearly outline the key challenges in multi-image understanding and present CcDPO as a two-level solution that directly addresses these issues. I describe core methods—structured captioning, perturbation-based training, and region-focused prompts—and introduce the MultiScope-42k dataset to support them. These contributions align closely with the claims, accurately reflecting the scope of my work.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss that CcDPO is designed for general multi-image reasoning and does not explicitly model temporally correlated inputs like videos, which may limit its performance in such scenarios. Additionally, due to limited OCR-focused data in our training set, the model may underperform on text-heavy tasks requiring fine-grained text recognition.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide detailed implementation settings, including model architectures, training stages, learning rates, loss formulations (Eq.1, Eq.2), optimization strategies (LoRA and full-parameter tuning), and hyperparameters (e.g., temperature β, loss weight γ). I also clearly state the number and type of GPUs used. The evaluation protocol is well-specified, listing all benchmarks and their assessed capabilities. These details are sufficient to reproduce the main experimental results and validate the core claims, even if code and data are not directly released.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We have provided open GitHub code and sufficient reproduction instructions.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We specify the training procedure, learning rates, optimizer settings (LoRA and full fine-tuning), key hyperparameters (β, γ), and hardware setup. Evaluation is conducted on standard benchmarks.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We report key results with sufficient experimental comparisons to support our main claims. While we do not provide detailed error bars in all tables or figures, the results are consistent across benchmarks and training setups, and variability was assessed during model development to ensure robustness.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We specify the number and type of GPUs used (8 GPUs with 90GB memory each).

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: All authors have reviewed and confirmed that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Our work improves the reliability of multi-modal large language models in multi-image reasoning tasks by reducing hallucinations and enhancing visual grounding. This has positive societal impact in applications requiring accurate visual understanding, such as education, medical imaging, and scientific analysis. However, more powerful image understanding capabilities may also raise risks related to surveillance, privacy invasion, or misuse in generating misleading content. We leave a deeper discussion of these impacts to future work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: Our work focuses on reducing hallucinations and improving grounding in multi-image understanding tasks, which poses relatively low risk of misuse. As such, we do not introduce high-risk components such as large-scale generative models or scraped dataset.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All benchmark datasets and compared methods are properly cited with appropriate references, and their licenses and use terms are respected.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We introduce a new dataset, MultiScope-42k, which is automatically generated and used to support our two-level preference optimization framework. The dataset construction process is thoroughly documented in the paper, including how chosen and rejected pairs are created at both the context and needle levels. We provide clear descriptions of the generation pipeline, perturbation strategies, and supervision signals, ensuring transparency and reproducibility.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper relies solely on existing publicly available benchmark datasets for both training and evaluation. It does not involve any crowdsourcing, user studies, or research with human subjects.

Guidelines:

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

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

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

    Answer: [NA]

    Justification: No human subjects were directly involved in this research as we used existing datasets.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: The core method in this research builds on existing multi-modal large language models (e.g., Qwen2-VL, LLaVA-OV) as backbones but does not introduce new LLM architectures or non-standard modifications. Our contribution lies in the training framework (CcDPO) and dataset construction, not in the development or adaptation of the LLMs themselves.

    Guidelines:

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