S2H-DPO: Hardness-Aware Preference Optimization for Vision–Language Models

Nitish Shukla[1] **Surgan Jandial**[2] **Arun Ross**[1]

[1]Michigan State University [2]Carnegie Mellon University

Correspondence: shuklan3@msu.edu

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable progress in single-image understanding, yet effective reasoning across multiple images remains challenging. We identify a critical capability gap in existing multi-image alignment approaches: current methods focus primarily on localized reasoning with pre-specified image indices ("Look at Image 3 and..."), bypassing the essential skills of global visual search and autonomous cross-image comparison. To address this limitation, we introduce a Simple-to-Hard (S2H) learning framework that systematically constructs multi-image preference data across three hierarchical reasoning levels requiring an increasing level of capabilities: (1) single-image localized reasoning, (2) multi-image localized comparison, and (3) global visual search. Unlike prior work that relies on model-specific attributes, such as hallucinations or attention heuristics, to generate preference pairs, our approach leverages prompt-driven complexity to create chosen/rejected pairs that are applicable across different models. Through extensive evaluations on LLaVA and Qwen-VL models, we show that our diverse multi-image reasoning data significantly enhances multi-image reasoning performance, yielding significant improvements over baseline methods across benchmarks. Importantly, our approach maintains strong single-image reasoning performance while simultaneously strengthening multi-image understanding capabilities, thus advancing the state of the art for holistic visual preference alignment.

1 Introduction

Vision–Language Models (VLMs) (Li et al., 2025; Bai et al., 2023; Lin et al., 2024) have made rapid progress, yet their performance drops for challenging tasks and inputs. In particular, we study reasoning of VLMs across multiple images where the model must not only interpret each image in isolation, but also align, compare, and integrate evidence across images. While proprietary systems such as GPT-4o (Hurst et al., 2024) demonstrate strong multi-image capabilities, open-source VLMs (Li et al., 2025; Zhang et al., 2024) still struggle to reliably aggregate information when visual context is distributed across several images. The difficulty compounds as the number of images grows, increasing both the search space and the need for consistent cross-image correspondence. Effective multi-image reasoning therefore hinges on two core abilities: (1) localizing where to look across multiple images, and (2) composing information from multiple regions into a coherent conclusion. Existing work primarily addresses this via multi-image pre-training (Awadalla et al., 2023) and supervised fine-tuning (SFT) (Jiang et al., 2024; Chen et al., 2024b; Liu et al., 2024b), with only a few studies (Liu et al., 2025) approaching the problem from the perspective of visual preference alignment. In particular, (Liu et al., 2025) observes that simply incorporating multi-image data during SFT is not sufficient to improve multi-image reasoning—and can, in some cases, even degrade performance on single-image tasks—thereby motivating Direct Preference Optimization (DPO)-style

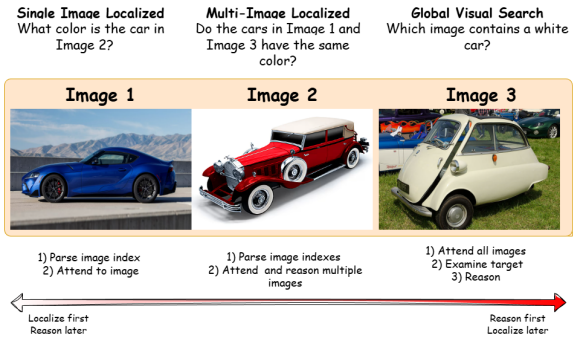

Figure 1: Multi-image reasoning skill hierarchy: From localized reasoning ($L1$: "What is the color of the car in Image 2?") to global visual search ($L3$: "Which image contains a white car?"). Each level requires strictly more capabilities than the previous.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36612–36623

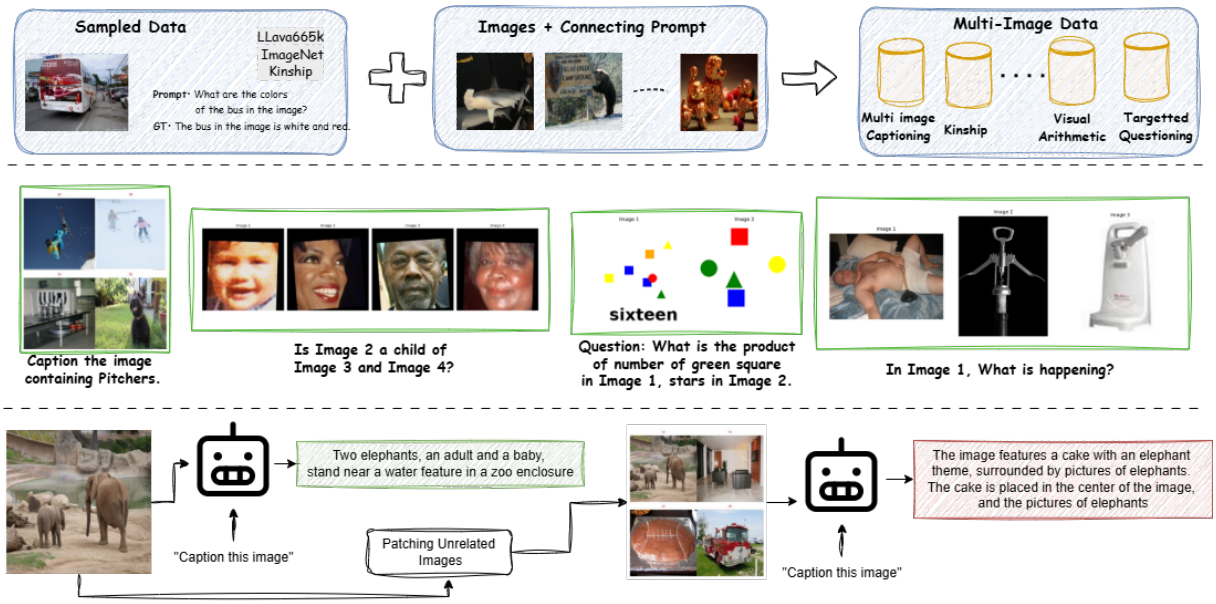

Figure 2: **Overview of our method.** We transform single-image data sources (e.g., LLaVA 665k, ImageNet) into multi-image datasets across visual modalities with progressively increasing cognitive load. Our synthetic transformation pipeline creates task hierarchies that advance from basic single-image understanding to complex multi-image reasoning requiring visual comparison, spatial reasoning, and cross-modal integration.

preference alignment for multi-image reasoning. However, a key challenge is constructing multi-image preference pairs because (1) each example must include a carefully selected set of multiple images and corresponding questions, and (2) each example requires well-curated chosen and rejected text response. These requirements make large-scale human annotation prohibitively expensive, motivating synthetic data approaches for generating multi-image preference data.

While recent work—most notably MIA-DPO (Liu et al., 2025)—proposes promising strategies for generating synthetic preference pairs, its data curation remains fundamentally limited in capturing the skills required for robust multi-image reasoning. In particular, it relies on questions that pre-specify where the model should look (e.g., "What is the color of the car in Image 2?" given multiple images). By explicitly indicating which image(s) to examine, this formulation sidesteps key competencies in which the model must autonomously determine relevant visual evidence, compare images, and compose information across multiple views. Therefore, we argue that, beyond simply reducing the cost of large-scale data generation, it is crucial to explicitly define the capabilities needed for multi-image reasoning and to ensure that data curation targets them. In this regard, let us consider the progression of multi-image reasoning capabilities:

- **Level 1 (Single-Image Localized):** "What is the color of the car in Image 2?" — Reason about one pre-specified image.
- **Level 2 (Multi-Image Localized):** "Do the cars in Image 1 and Image 3 have same color?" — Reason and Compare across multiple pre-specified images, unlike Level 1 which reduces to single-image reasoning.
- **Level 3 (Global Visual Search):** "Which images contain a white car?" — Reason and Search all images to locate relevant regions and carefully aggregate information, unlike Level 1 and Level 2, where regions are pre-specified.

Figure 1 illustrates these capabilities, where each level demands strictly more from the model than the previous level. Critically, MIA-DPO (Liu et al., 2025) trains only on Level 1. We argue this is not merely a difference in question format, but an important gap: different formulations induce qualitatively different reasoning patterns with increasing difficulty. In particular, at Level 1 the model reasons about a single, pre-specified image ("look at Image 2"); at Level 2 it must compare multiple pre-specified images ("look at Image 1 and Image 3"); and at Level 3 it must autonomously determine where to look and then reason over the relevant regions (e.g., "look for the image with a white car"). Thus, we ask the question: **Can preference tuning**

on diverse multi-image reasoning tasks achieve broader generalization than index-based localization training alone?

More precisely, we automatically construct a diverse synthetic training set containing preference pairs spanning Levels 1-3, allowing models to learn multi-image reasoning capabilities simultaneously. We then perform a **comprehensive multi-image DPO approach that jointly trains models across all reasoning levels**. Through systematic ablations, we demonstrate that: (1) joint training on mixed-difficulty data enables robust multi-image reasoning, and (2) models trained on diverse question types, including global search, generalize better than those trained only on localized questions.

Experimental results highlight a significant capability gap between our approach and the current state-of-the-art. While MIA-DPO demonstrates moderate accuracy on lower-complexity benchmarks, its performance notably declines when faced with more challenging Level 3 tasks. Conversely, our method—which leverages training samples across all reasoning levels—consistently outperforms MIA-DPO on both tiers, with a more pronounced advantage on the harder tasks (see Table 1). A critical limitation of MIA-DPO is its reliance on Level 1 examples, which hinge on model-specific intrinsic hallucinations to generate chosen-rejected pairs. This necessitates the generation of a new dataset for every new model architecture. By introducing higher hardness levels through prompt-driven complexity, our method alleviates this dependency. Furthermore, our approach preserves single-image reasoning capabilities, matching the performance of the pre-finetuned baseline. These findings demonstrate that our method substantially enhances multi-image reasoning across all semantic levels without compromising foundational single-image performance.

In summary, our main contributions are:

- We identify a "capability gap" in existing multi-image alignment: current methods train only on localized reasoning with pre-specified images (Level 1), missing the global and partial visual search capabilities (Levels 2-3) required for real-world multi-image understanding.
- We develop a systematic technique for generating high-quality rejected answers from correct-answer-only datasets, enabling comprehensive multi-image preference optimization across all reasoning levels without manual annotation.
- Through extensive experiments, we demonstrate that our comprehensive multi-image DPO approach substantially outperforms existing methods while preserving single-image reasoning capabilities.

2 Proposed Method

2.1 Preliminaries

In this section, we present the concept of visual preference alignment and use the Direct Preference Optimization (DPO) method as a representative example.

Visual Preference Alignment: Preference alignment focuses on aligning a model's outputs with a set of preferences. Common methods include **R**einforcement **L**earning from **H**uman **F**eedback (**RLHF**) (Yu et al., 2024a) and **R**einforcement **L**earning from **AI F**eedback (**RLAIF**)(Yu et al., 2024b). Consider a dataset D where each sample consists of an input prompt x, a preferred response y_w, and a rejected response y_l. Formally, we represent the dataset as $D = \{x, y_w, y_l\}$. The input x can be an interleaved sequence of images v and text t.

When a VLM processes x to generate an output y, a reward $r(x, y)$ is assigned by a reward model r, which assigns higher scores to preferred outputs and lower scores to rejected ones. Visual preference alignment aims to maximize this reward:

$\max_\theta \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}\left[r(x, y)\right]$, where θ, π_θ and $\pi_\theta(y|x)$ denote the parameters, policy, and output distribution of VLM, respectively.

To avoid overfitting on the dataset D, a KL-divergence loss D_{KL} is incorporated by preference alignment approaches to regularize the difference between the model's policy $\pi_\theta(y|x)$ and a reference model's policy $\pi_{\text{ref}}(y|x)$:

$$\max_\theta \left[\mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}\left[r(x, y)\right] - \beta \cdot D_{\text{KL}}(\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x))\right], \quad (1)$$

where, the hyper-parameter β controls the influence of KL-divergence on the optimization objective. Note that the reference model is the model's state prior to preference alignment.

Direct Preference Optimization (DPO): To optimize the preference alignment objective in Eq. (1), we can use either an online reward model (*e.g.*, PPO (Schulman et al., 2017)) or pre-computed off-line chosen/rejected pairs (*e.g.*, DPO (Rafailov et al., 2024)). Given its simplicity, DPO has been

Figure 3: Simple-to-Hard (S2H) DPO Data Format. **Level 1:** The query directly points to a specific image (*e.g.*, "Caption the first image"), and rejected pairs provide responses completely unrelated to the target image. **Level 2:** The query references multiple images to enforce multi-image reasoning (*e.g.*, "Compare the first and third images"). **Level 3:** The query is open-ended and requires the model to examine all images before identifying which one satisfies the semantic constraint (*e.g.*, "Caption the image containing a peacock").

widely adopted in previous visual alignment works (Zhao et al., 2023; Zhou et al., 2024). Eq. (1) can be reformulated as the loss function of DPO: $L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$

$$-\mathbb{E}_{(x,y_w,y_l)\sim D}\Big[\log\sigma\big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big)\Big], \tag{2}$$

where, $\sigma(.)$ denotes the sigmoid function. As shown in Eq. (2), DPO-based alignment methods concentrate on designing input prompts x and pairing each with a preferred output y_w and a rejected output y_l.

2.2 Simple-to-hard (S2H) Tasks

We propose three Simple-to-Hard (S2H) probe tasks with increasing levels of complexity, defined by the cognitive load required for localization, cross-image grounding, and reasoning. Our tasks therefore escalate from basic single-image query to complex multi-step reasoning that demands joint understanding of all images. We illustrate the overall process in Figure 2. Below, we explain these tasks in detail.

2.2.1 Single Image Localized (Level 1)

In the simplest case, following prior work (Liu et al., 2025), we convert single-image datasets to multi-image format by appending unrelated distractor images to existing VQA samples. The chosen response corresponds to the gold answer for the target image, while the rejected response is an incorrect answer generated by the model. Critically, this approach does not establish explicit relationships between images—the task simply requires the model to ignore irrelevant visual context. The rejected responses primarily arise from model hallucinations when processing noisy multi-image inputs, rather than from meaningful cross-image reasoning errors. While this provides a baseline for multi-image handling, it does not test genuine multi-image reasoning capabilities.

2.2.2 Multi-Image Localized (Level 2)

In this category, we extend Level 1 tasks to a multi-image setting. Rather than analyzing a single image in isolation, we expand the context by requiring the model to localize features across multiple images and respond to prompts that necessitate explicit cross-image connections. We introduce two distinct tasks in this category: (i) Kinship Recognition, where the objective is to determine familial or social relationships within a set of facial images; and (ii) Visual Arithmetic, which requires the model to perform mathematical reasoning—such as counting or logical operations—on shapes, numbers, and objects distributed across several images. We select Kinship Recognition and Visual Arithmetic as Level-2 tasks because they capture complementary

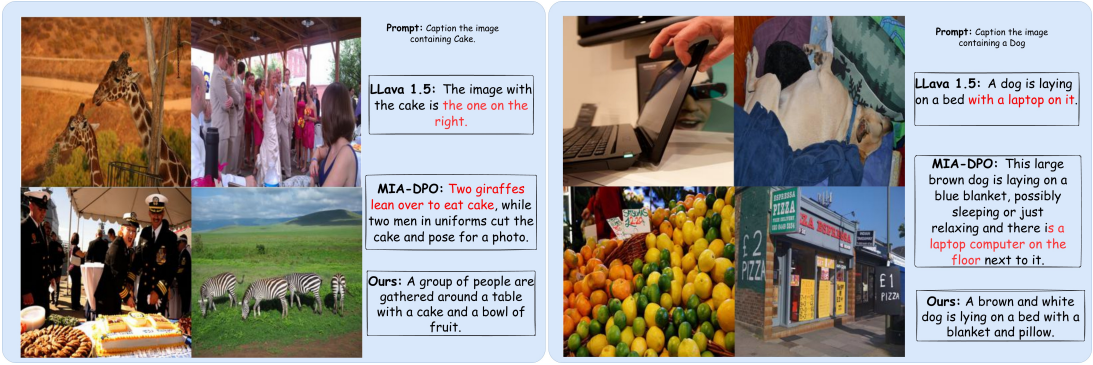

Figure 4: Outputs of our method on Global Visual Search. Our approach effectively isolates the queried concept and produces accurate responses to the prompt.

aspects of multi-image localized reasoning. Kinship Recognition requires asymmetric relational inference: although multiple images are presented, the answer pertains primarily on a single target image in relation to others, demanding cross-image relational grounding. Visual Arithmetic, in contrast, requires symmetric compositional aggregation, where each image contributes equally and the model must combine information across all images to compute the answer. Together, these sub-skills capture the core challenges of L2 multi-image reasoning.

Chosen/Rejected pair generation. For Kinship Recognition, we leverage preexisting datasets (Wang et al., 2017; Robinson et al., 2016) to generate chosen and rejected pairs. The process is straightforward: we first read the labels and relationships of the individuals. If the relationship is correct, we generate a chosen deterministic caption (e.g., 'Yes, Image 3 is…') and a corresponding rejected caption (e.g., 'No, Image 3 is not…'). Conversely, if the relationship is incorrect, we generate a chosen–rejected pair by flipping the responses accordingly. For Visual Arithmetic, images are generated synthetically using deterministic procedures, ensuring that each image contributes clearly to the final answer while maintaining controlled, reproducible inputs for multi-image reasoning. Images are generated synthetically by placing random shapes and objects in each image while keeping track of their counts. To create a question, we randomly sample a subset of images, select some objects, and choose a random mathematical operator (e.g., addition, subtraction) to form a problem whose answer depends on aggregating information across the selected images. This deterministic procedure guarantees correct chosen/reject pair generation.

2.2.3 Global Visual Search (Level 3)

This category inverts the computational ordering: the model must first perform global reasoning across the entire image set before localizing the relevant image. Consider the task of captioning an image that satisfies a semantic criterion, *e.g.*, "Caption the image containing a peacock". Unlike previous levels, localization cannot precede reasoning. Instead, the model must evaluate all images against the query constraint, identify the matching instance, and only then generate the caption. This formulation tests the model's ability to integrate cross-image search with subsequent generative tasks.

Single-to-Multi-Image Prompt Construction. Figure 3 (bottom) illustrates our procedure for converting single-image captions into multi-image chosen-rejected pairs. We construct training data from existing annotated datasets through a two-stage process designed to distinguish targeted search from general summarization. *Chosen pair generation:* We first select a target concept from ImageNet (Deng et al., 2009) and sample a target image $\mathcal{I}_{\text{target}}$ belonging to that concept. We then prompt a vision-language model to generate both (1) a detailed caption and (2) the primary object present in the image (*e.g.*, a peacock). To form a multi-image set, we pair $\mathcal{I}_{\text{target}}$ with $N-1$ distractor images randomly sampled from *different* concept classes, ensuring that the same primary object does not appear in more than one image. The chosen response is then assigned as the generated caption. *Rejected pair generation:* For the same multi-image set, we prompt the captioning model *without* specifying a target object, asking it to simply "caption the images." This yields a rejected response that typically either aggregates informa-

tion across multiple images or provides a generic description, failing to isolate the target. Importantly, these rejected responses are not randomly unrelated—they represent plausible but incorrect responses, which provides a better learning signal to model compared to trivial negatives. We use Qwen2.5-VL-32B (Bai et al., 2025b) as the caption generator in this work. *Quality filtering:* To ensure meaningful contrast between chosen and rejected pairs, we compute semantic similarity between the two responses using text encoders (CLIP, MPNet). Pairs with similarity exceeding threshold τ (top quartile) are discarded, as high similarity indicates the rejected response inadvertently captured the correct behavior. This filtering ensures the contrastive pairs provide a clear learning signal for distinguishing conditional visual search from unconstrained multi-image description.

2.3 Implementation Details

Our S2H-DPO is model-agnostic and can be directly applied to a wide range of LVLMs without requiring the generation of additional training datasets, in contrast to prior approaches. We evaluate our method on three widely used MLLMs: LLaVA-v1.5-7B (Liu et al., 2024a) and Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-2B (Bai et al., 2025a). All models are trained for three epochs with a learning rate of 5×10^{-5}, and the temperature parameter β is set to 0.1. We create 20K samples for each level.

3 Related Works in Preference Alignment

Aligning Large Language Models with human preferences has become a critical requirement for their safe and reliable deployment in real-world applications (Betley et al., 2025; Turpin et al., 2023; Vidgen et al., 2024; Guan et al., 2025; Furniturewala et al., 2024), particularly under stringent ethical and safety constraints. Preference alignment methods can be broadly divided into two categories. The first involves feedback-driven alignment, which leverages either human-annotated preferences (Rafailov et al., 2024; Bai et al., 2022) or AI-generated feedback (Yu et al., 2024b). The second category focuses on prompt-based guidance, where carefully designed instructions are used to steer model behavior without modifying model parameters (Chen et al., 2023).

Similarly, recent works investigate the unreliabilty and misaligned behavior of vision–language large models (VLLMs) across various tasks (Jandial et al., 2025, 2026; Das et al., 2026; Rahmanzadehgervi et al., 2025), motivating the development of specialized vision-language preference alignment approaches. (Sun et al., 2023) introduced LLaVA-RLHF, which utilizes human-annotated preference data to reduce hallucinations in LLaVA. (Li et al., 2024) proposed a preference distillation framework that transfers alignment signals into VLLMs, improving visual grounding and response relevance. Similarly, (Yu et al., 2024a) collected fine-grained human preferences in the form of segment-level corrections to hallucinated content and optimized model behavior using dense supervision. HA-DPO (Zhao et al., 2023) addresses alignment by leveraging GPT-4's API to create the required DPO data, though it incurs substantial API expenses. POVID (Zhou et al., 2024) induces hallucinations using blurred images GPT-4 to generate DPO data. Recently, MIA-DPO (Liu et al., 2025) automates the generation of chosen-rejected pairs by exploiting the model's intrinsic hallucinations, but it still necessitates creating a fresh set of DPO data for each model.

4 Experiments

4.1 Benchmarks

To assess our method, we evaluate its performance across two primary categories of benchmarks. First, we examine multi-image reasoning using three multi-image benchmarks BLINK (Fu et al., 2024), MANTIS (Jiang et al., 2024) and NLVR2 (Suhr et al., 2019). Second, to ensure our approach maintains robust single-image capabilities, we test on two single-image benchmarks, including MM-Star (Chen et al., 2024a) and POPE (Li et al., 2023). This extensive evaluation demonstrates our method's versatility and confirms significant performance gains in multi-image scenarios without sacrificing single-image proficiency.

4.2 Baseline Methods

We benchmark our method against several state-of-the-art preference optimization strategies. LLaVA-RLHF (Sun et al., 2023) serves as the standard RLHF baseline, utilizing human and GPT-generated feedback. HA-DPO (Zhao et al., 2023) focuses on error correction by leveraging GPT-4V as an expert annotator to rectify model hallucinations. In contrast, POVID (Zhou et al., 2024) adopts an adversarial approach, pairing distorted

images with synthetic hallucinations to stress-test modality alignment. Finally, we compare against MIA-DPO (Liu et al., 2025), which automates the generation of preference data by exploiting the model's hallucinations in multi-image contexts.

Table 1: **Main results on multi-image benchmarks.** We compare our S2H-DPO with other DPO algorithms across five multi-image benchmarks. Our method yields consistent improvements over both the classic LLaVA-v1.5 and the recent Qwen2.5-VL-7B. In contrast, single-image DPO methods perform poorly on multi-image benchmarks.

Models	Parameter	BLINK	MANTIS	NLVR2	Average
GPT-4V (Achiam et al., 2023)	-	51.1	62.7	88.8	67.53
LLaVA-v1.6 (Li et al., 2025)	7B	39.6	45.6	58.9	48.03
Qwen-VL-Chat (Bai et al., 2023)	7B	31.2	39.2	58.7	43.03
VideoLLaVA (Lin et al., 2024)	7B	38.9	35.9	56.5	43.77
Fuyu (Bavishi et al., 2023)	8B	36.6	27.2	51.1	38.3
Idefics2 (Laurençon et al., 2024)	8B	45.2	48.9	86.9	60.33
InstructBLIP (Dai et al., 2023)	13B	42.2	45.6	60.3	49.37
CogVLM (Wang et al., 2024)	17B	41.5	45.2	58.6	48.43
Emu2-Chat (Sun et al., 2024)	37B	36.2	37.8	58.2	44.07
LLaVA-v1.5 (Liu et al., 2024a)	7B	37.1	41.9	52.1	43.7
+ LLaVA-RLHF (Sun et al., 2023)	7B	40.8	30.4	51.8	41.0
+ HA-DPO (Zhao et al., 2023)	7B	38.6	34.6	51.6	41.6
+ POVID (Zhou et al., 2024)	7B	19.9	37.8	21.4	26.37
+ MIA-DPO (Liu et al., 2025)	7B	42.9	44.2	54.2	47.1
+ S2H-DPO (Ours)	7B	**43.40**	**47.93**	**55.59**	**48.97**
Δ	-	**+6.30**	**+6.03**	**+3.49**	**+5.27**
Qwen2.5-VL (Bai et al., 2025b)	7B	54.29	68.66	74.28	65.74
+ MIA-DPO (Liu et al., 2025)	7B	41.28	59.45	74.18	58.30
+ S2H-DPO (Ours)	7B	**55.85**	**74.19**	**74.67**	**68.24**
Δ	-	**+1.56**	**+5.53**	**+0.39**	**+2.49**
Qwen3-VL (Bai et al., 2025a)	2B	51.61	79.61	49.71	60.31
+ S2H-DPO (Ours)	2B	**53.92**	**81.71**	**50.61**	**62.08**
Δ	-	**+2.31**	**+2.10**	**+0.90**	**+1.77**

4.3 Results on Multi-Image Benchmarks

Results on LLaVA-1.5: We report the performance of our method across several multi-image benchmarks in Table 1. Our approach yields significant gains, specifically improving by 6.30%, 6.03%, and 3.49% across three key datasets. Notably, on the complex BLINK benchmark—which requires specialized domain knowledge—S2H-DPO outperforms the LLaVA-v1.5 baseline by 6.30%. Furthermore, evaluations on the MANTIS and NLVR2 benchmarks show improvements of 6.03% and 3.49%, respectively. These results clearly highlight the effectiveness of our proposed method in enhancing the model's capacity for visual synthesis and reasoning within multi-image contexts. We illustrate these results in Figure 4.

Comparison with Preference Optimization baselines: As shown in Table 1, S2H-DPO consistently outperforms established preference optimization methods. While standard LLaVA-RLHF and HA-DPO show marginal or even negative gains on specific multi-image tasks, our method achieves substantial improvements across all benchmarks.

Table 2: **Main results on single-image benchmarks.** We compare S2H-DPO with other DPO approaches across two key single-image benchmarks. S2H-DPO maintains strong proficiency in single-image tasks.

Models	Parameter	MMStar	POPE	Average
LLaVA-v1.6 (Li et al., 2025)	7B	37.6	70.3	53.95
Qwen-VL-Chat (Bai et al., 2023)	7B	34.5	74.9	54.7
Idefics2 (Laurençon et al., 2024)	8B	49.5	86.2	67.85
OpenFlamingo (Awadalla et al., 2023)	9B	36.9	52.6	44.75
InstructBLIP (Dai et al., 2023)	13B	32.7	86.1	59.4
CogVLM (Wang et al., 2024)	17B	39.9	88.0	63.95
Emu2-Chat (Sun et al., 2024)	37B	40.7	88.0	64.35
LLaVA-v1.5 (Liu et al., 2024a)	7B	32.9	85.9	59.4
+ LLaVA-RLHF (Sun et al., 2023)	7B	31.6	80.8	56.2
+ HA-DPO (Zhao et al., 2023)	7B	33.5	84.3	58.9
+ POVID (Zhou et al., 2024)	7B	36.2	86.3	**61.25**
+ MIA-DPO (Liu et al., 2025)	7B	32.9	**87.2**	60.05
+ S2H-DPO (Ours)	7B	**33.62**	85.70	59.66
Qwen2.5-VL (Bai et al., 2025b)	7B	62.24	83.13	72.69
+ S2H-DPO (Ours)	7B	**62.47**	**83.59**	**73.03**
Qwen3-VL (Bai et al., 2025a)	2B	53.42	85.28	69.35
+ S2H-DPO (Ours)	2B	**53.62**	**85.46**	**69.54**

Notably, compared to the closest baseline, MIA-DPO, our approach yields an additional boost of 0.50%, 2.34%, 1.39%. These results indicate that our method provides a more robust signal for multi-image alignment than existing DPO or RLHF variants.

Results on Qwen2.5-VL-7B: As shown in Table 1, our method consistently improves Qwen2.5-VL-7B across all three multi-image benchmarks, yielding gains of 1.56%, 5.53%, and 0.39%, with an average improvement of 2.49%. In contrast to preference-optimization baselines such as MIA-DPO, which degrades performance on the first two benchmarks, S2H-DPO achieves stable and substantial improvements across all settings, demonstrating its effectiveness for multi-image alignment and reasoning.

Results on Qwen3-VL-2B: To further validate the generalizability of S2H-DPO, we apply it to the recently released Qwen3-VL-2B (Bai et al., 2025a), a compact 2B-parameter model. As shown in Tables 1 and 2, S2H-DPO consistently improves performance across both multi- and single-image benchmarks. On multi-image benchmarks, S2H-DPO delivers consistent gains across all three evaluations, improving BLINK by +2.31, MANTIS by +2.10, and NLVR2 by +0.90, resulting in an overall average gain of +1.77 (60.31 → 62.08). On single-image tasks, our method also achieves steady gains on both MMStar (53.42 → 53.62) and POPE (85.28 → 85.46), yielding an average improvement of +0.19. These results demonstrate that S2H-DPO scales effectively to smaller, more recent architectures, achieving meaningful multi-

image alignment improvements without sacrificing single-image proficiency.

4.4 Results on Single-Image Benchmarks

While S2H-DPO is primarily designed to enhance performance in multi-image reasoning tasks, it is crucial that the optimized model maintains effectiveness on its core capability: single-image reasoning. We evaluate S2H-DPO on established single-image benchmarks to verify this property. As shown in Table 2, S2H-DPO consistently outperforms the LLaVA-v1.5 baseline and existing preference-alignment methods, including LLaVA-RLHF and HA-DPO, across all evaluated single-image benchmarks. These results demonstrate that S2H-DPO not only substantially improves multi-image reasoning performance but also preserves strong single-image capabilities. This dual proficiency makes S2H-DPO well-suited for training robust vision-language models capable of handling both single- and multi-image scenarios in real-world deployments.

5 Ablation Studies

Ablation on curriculum training and data mixture: We study the effect of curriculum learning strategies and data mixture on multimodal reasoning performance. Specifically, we compare incremental curriculum-based training—where models are exposed to tasks in increasing order of complexity—against flat training baselines that directly optimize on the target task distribution using different data mixtures. As summarized in Table 3, we denote incremental setups as $A \rightarrow B$, indicating training on task set B after pretraining on A. Across all mixtures, flat training consistently outperforms curriculum-based approaches. For example, L2 flat training attains a mean accuracy of 48.13%, 3.3% improvement over the L1→L2 curriculum (44.83%). This improvement corresponds to gains of 3.9% and 5.1% on BLINK and MANTIS, respectively. This trend generalizes across task combinations: L3 flat training improves upon L1→L3 by 3.0% (47.27% vs. 44.24%), while (L2∪L3) flat training yields a 1.0% accuracy increase over L1→(L2∪L3) (46.53% vs. 45.56%). Even the most gradual curriculum, L1→L2→L3, achieves only 45.55% mean accuracy, underperforming all flat training variants. **These results suggest that, for multimodal reasoning tasks, direct exposure to the target task distribution is more effective than gradual adaptation through curriculum hierarchies**. This is because curricula based training on simpler tasks induce myopic reasoning behavior, causing models to over-rely on localized or explicitly referenced visual cues present in the prompt, rather than developing a holistic, global understanding of the visual scene.

Table 3: Effect of curriculum training strategies. Comparison of simple-to-hard (S2H) curriculum training against flat training. Flat training outperforms incremental curriculum training, as the latter leads to myopic optimization that biases the model to look for localization cues in the prompt.

Curriculum	BLINK	MANTIS	NLVR2	Mean
LLaVA-v1.5	37.1	41.9	52.1	43.70
+L1 (Liu et al., 2025)	42.9	44.2	54.2	47.10
L1→ L2	40.2	39.6	54.7	44.83
L2 flat	44.1	44.7	55.6	48.13
L1→ L3	37.03	41.0	54.7	44.24
L3 flat	43.8	42.8	55.2	47.27
L1→ (L2∪L3)	40.17	41.0	55.5	45.56
(L2∪L3) flat	43.2	41.4	55.0	46.53
L1→ L2→ L3	39.96	41.4	55.3	45.55

Table 4: DPO vs. SFT on LLaVA-v1.5. Preference-based optimization (DPO) provides stronger guidance than standard supervised fine-tuning (SFT), leading to improved performance across benchmarks.

Training	BLINK	MANTIS	NLVR2	Mean
LLaVA	37.1	41.9	52.1	43.70
+SFT	40.52	45.15	55.09	46.92
+DPO	43.40	47.93	55.59	48.97

Ablation on SFT vs DPO: We compare the effectiveness of supervised fine-tuning (SFT) and direct preference optimization (DPO) for enhancing multimodal reasoning capabilities. We simply convert the DPO data to SFT data by discarding the rejected samples. As shown in Table 4, both approaches substantially improve over the LLaVA-v1.5 baseline (43.70% mean accuracy), with SFT achieving 46.92% and DPO reaching 48.97% mean accuracy. DPO demonstrates consistent advantages across all benchmarks, outperforming SFT by 2.88% on BLINK (43.40% vs. 40.52%), 2.78% on MANTIS (47.93% vs. 45.15%), and 0.41% on NLVR2 (55.59% vs. 55.09%). These results indicate that preference-based optimization provides stronger learning signals for multimodal reasoning tasks compared to standard supervised learning.

Ablation on number of distractors. We study the effect of distractor count in the L3 task by training separate models on 20K samples with 2–5 dis-

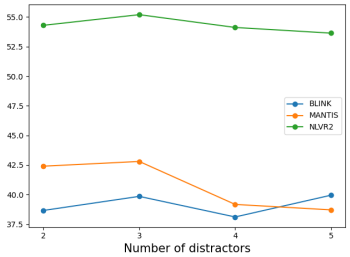

Figure 5: Effect of the number of distractors (2–5) on L3 task performance across BLINK, MANTIS, and NLVR2, showing peak accuracy at 3 distractors.

tractors. Figure 5 shows the resulting accuracy trends. Performance improves as distractors increase, peaks at 3, and declines beyond that. This suggests that while more distractors enrich supervision and encourage robust reasoning, too many introduce noise. Overall, 3 distractors provide the optimal balance between contextual diversity and reasoning difficulty across benchmarks.

Does S2H training enhance L3 ability? To evaluate this, we construct L3-style questions using the ImageNet validation set. For each question, we randomly sample a target image belonging to a object c, and select $N \in \{2, 3, 4\}$ distractor images from different classes. The model is then asked: "Which of the following images contains `object`?" On 500 such questions, our fine-tuned model achieves an accuracy of **31.05%**, outperforming baseline LLaVA-v1.5 by **6.94%**. This improvement indicates that the proposed S2H training tasks enhances higher-order visual reasoning and leads to stronger L3 capability.

Table 5: Comparison of DPO and GRPO fine-tuning on Qwen3-VL-2B across multi-image reasoning benchmarks. While GRPO achieves the highest overall performance, it exhibits instability across benchmarks.

Method	MANTIS	BLINK	NLVR2
Qwen3-VL-2B	47.00	79.61	49.71
+ DPO	53.92	81.71	50.61
+ GRPO	61.29	70.62	61.73

Effects of on-policy optimization To further validate the effectiveness of our data generation strategy, we conduct additional experiments using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as an on-policy reinforcement learning alternative to DPO. We evaluate three reward functions on Qwen3-VL-2B: (i) a *length-based* reward that encourages responses close to an optimal length of 20 tokens; (ii) a *format-based* reward that enforces adherence to a predefined output structure; and (iii) a *rule-based* reward that measures similarity between the generated output and the gold reference answer.

As shown in Table 5, GRPO yields substantial gains over DPO on MANTIS and NLVR2, but exhibits a performance regression on BLINK. We attribute this instability to GRPO's sensitivity to reward design and optimization hyperparameters, a challenge orthogonal to the primary focus of this work. Nonetheless, the strong improvements achieved with our simple reward setup further corroborate the quality of our data generation strategy for multi-image reasoning.

6 Conclusion

We identify and address a critical capability gap in multi-image reasoning within Vision-Language Models. While prior methods focused heavily on localized indexing, our Simple-to-Hard (S2H) framework demonstrates that multi-image reasoning requires a hierarchical approach—advancing from basic isolation to complex global visual search and cross-image composition. By systematically generating synthetic preference pairs across these cognitive levels, we move beyond model-specific hallucinations toward a more robust, prompt-driven alignment strategy. Our results show that this comprehensive training not only significantly improves performance on complex multi-image benchmarks but also preserves the foundational single-image reasoning capabilities.

7 Limitations

Despite its effectiveness, our S2H-DPO framework has few limitations. First, although preference pairs are generated without human annotation, the process still relies on existing datasets and a strong captioning model, whose biases may propagate into the training data. Second, our experiments focus on multi-image settings with a limited number of images (capped at 6 images); scaling to much larger image sets may introduce additional challenges in attention and efficiency that we do not explore. Finally, while S2H-DPO captures key forms of multi-image reasoning, it does not cover all possible reasoning patterns, such as temporal or causal reasoning across images, which we leave for future work.

References

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

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Tasırlar. 2023. Introducing our multimodal models. *Adept Blog*.

Jan Betley, Daniel Chee Hian Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 4043–4068.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024b. ShareGPT4Video: improving video understanding and generation with better captions. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*.

Zhipeng Chen, Kun Zhou, Zheng Zhang, Beichen Gong, Xin Zhao, and Ji-Rong Wen. 2023. ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. 2026. More Images, More Problems? A Controlled Analysis of VLM Failure Modes. *arXiv preprint arXiv:2601.07812*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. 2025. The order effect: Investigating prompt sensitivity to input order in llms. *arXiv preprint arXiv:2502.04134*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Surgan Jandial, Yinheng Li, Justin Wagle, and Kazuhito Koishida. 2026. Do GUI grounders truly understand UI elements? In *Proceedings of Findings of the Association for Computational Linguistics: EACL*.

Surgan Jandial, Yinong Oliver Wang, Andrea Bajcsy, and Fernando De la Torre. 2025. On the fine-grained planning abilities of VLM web agents. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? In *Proceedings of The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. 2025. LLaVA-neXT-interleave: Tackling multi-image, video, and 3D in large multimodal models. In *Proceedings of The Thirteenth International Conference on Learning Representations (ICLR)*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024. VLFeedback: A large-scale AI feedback dataset for large vision-language models alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of The Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, and 1 others. 2024b. MMDU: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for LVLMs. *arXiv preprint arXiv:2406.11833*.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2025. MIA-DPO: Multi-image augmented direct preference optimization for large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2025. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*.

Joseph P. Robinson, Ming Shao, Yue Wu, and Yun Fu. 2016. Families in the wild (fiw): Large-scale kinship image database and benchmarks. In *Proceedings of the ACM on Multimedia Conference*, pages 242–246.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Bertie Vidgen, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Max Bartolo, Borhane Blili-Hamelin, Kurt Bollacker, Rishi Bomassani, Marisa Ferrara Boston, Siméon Campos, Kal Chakra, Canyu Chen, Cody Coleman, Zacharie Delpierre Coudert, and 81 others. 2024. Introducing v0.5 of the AI Safety Benchmark from MLCommons. *arXiv preprint arXiv:2404.12241*.

Shuyang Wang, Joseph P Robinson, and Yun Fu. 2017. Kinship verification on families in the wild with marginalized denoising metric learning. In *Proceedings of Automatic Face and Gesture Recognition (FG)*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. CogVLM: Visual expert for pretrained language models. In *Proceedings of The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024a. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024b. RlHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2405.17220*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, and 1 others. 2024. Internlm-Xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

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