

A Related Work

Existing approaches that apply DPO to align MLLMs fundamentally involves two components: (1) preference data construction and (2) optimization strategy. We accordingly present current multimodal DPO research from these two perspectives.

A.1 Preference Data Construction

Existing preference data construction approaches primarily fall into two categories: (1) comparative ranking-based methods and (2) hallucination correction-based methods.

The first category constructs preference data by directly evaluating the rankings of candidate responses. For example, CLIP-DPO [39] uses CLIP’s image-text similarity metrics to rank candidate responses and selects pairs with large ranking gaps and similar lengths as preference data. Beyond this, AMP [12] leverages MLLMs of varying scales to generate multi-level preference data, assuming that responses from larger MLLMs should rank higher. Furthermore, BPO [40] distorts the input image and pairs the response generated from the distorted image (ranked lower) with the response from the original image (ranked higher) to create preference pairs. To eliminate confounding factors such as text style that hinder the model’s ability to discern genuine trustworthiness differences within response pairs, RLAIIF-V [11] introduces a deconfounded candidate response generation strategy. This strategy generates candidate responses through multiple sampling decoding trials with different random seeds while keeping the input prompt and decoding parameters constant.

The second category constructs preference data by detecting and correcting hallucinated content in MLLM responses. For example, RLHF-V [10] relies on human annotators to identify and rectify hallucinatory content, ensuring the creation of higher-quality preference datasets. In contrast, HA-DPO [13] bypasses human annotators by leveraging GPT-4 [1] to detect and correct hallucination using fine-grained visual annotations from the Visual Genome dataset [41]. Similarly, OPA-DPO [19] employs GPT-4V [4] to rectify hallucination but identifies them through direct image pair comparisons, eliminating the need for fine-grained annotations and offering greater flexibility. To reduce the high API costs associated with commercial LLMs, HSA-DPO [37] trains a specialized hallucination detection model using a sentence-level annotation dataset generated by GPT. Moreover, building on RLAIIF-V [11], TPO [23] introduces a topic-level self-correction paradigm. This method works on identifying hallucination at the topic level through sub-sentence clustering, constructing topic-level preference pairs, and generating response preference pairs using a deconfounded topic-overwriting strategy—ensuring linguistic style consistency.

A.2 Optimization Strategies

Apart from enhancing preference data construction, researchers have also explored various optimization strategies to improve MLLM preference alignment. For example, MPO [42] introduces a mixed preference optimization model, which effectively combines preference optimization techniques and conventional supervised fine-tuning. To mitigate the hallucination of MLLMs, AMP [12] introduces a multi-level direct preference optimization algorithm, enabling robust multi-level preference learning, while CHiP [18] introduces a cross-modal hierarchical DPO model involving two key optimizations: hierarchical textual preference optimization for capturing fine-grained textual preferences and visual preference optimization for cross-modal preference alignment. To address the gradient vanishing problem induced by off-policy data, OPA-DPO [19] proposes an adaptive mechanism that dynamically balances exploration and exploitation during learning. Furthermore, MIA-DPO [43] specifically targets hallucination reduction in multi-image scenarios, where MLLMs process multiple input images simultaneously through optimized cross-image attention mechanisms.

One key issue in improving MLLMs’ reasoning ability is mitigating their over-reliance on textual prompts while enhancing visual content utilization. To address this issue, SymDPO [44] introduces a symbol demonstration direct preference optimization model for in-context learning, which strengthens MLLMs’ visual understanding by replacing textual answers with random symbols, thereby forcing MLLMs to establish mappings between visual information and symbolic responses. For more general multimodal understanding contexts, several studies [15, 16, 17, 18, 19] focused on extending DPO with vision-oriented preference learning to improve MLLMs’ visual signal interpretation. These approaches preserve DPO’s structural framework while only introducing visual variation in the

516 contrastive image-prompt-response triplet pairs. For contrastive image generation, existing methods
 517 employ diverse strategies: mDPO [15] applies geometric transformations to original images, V-
 518 DPO [16] employs a generative model to replace key visual elements through image inpainting [45],
 519 MFPO [17], OPA-DPO [19] and CHiP [18] perform noise injection to original images, while
 520 CHiP [18] utilizes a forward diffusion process [46] for generating contrastive images.

521 Although current vision-enhanced multimodal preference alignment methods have demonstrated great
 522 progress in reducing MLLM hallucination, they exhibit two critical limitations: (1) their DPO-based
 523 objective function derivations lack mathematical rigor, as they overlook the fact that the intractable
 524 partition functions in multimodal scenarios with different vision inputs cannot be directly canceled;
 525 and (2) they fail to provide direct preference supervision for DPO-based visual understanding en-
 526 hancement. To address these limitations, we propose Symmetric Multimodal Preference Optimization,
 527 which effectively utilize the corresponding preferred responses of contrastive images for optimizing
 528 the visual understanding capabilities of MLLMs.

529 B Impact of Partition Function in Multimodal Preference Optimization

530 Due to space limitations, Section 1 briefly argues that existing multimodal DPO methods [15, 16,
 531 17, 18, 19] non-rigorously ignore two partition functions in their vision-oriented contrastive learning
 532 mechanisms. Here, we present a detailed analysis of these functions’ roles in model optimization.

533 Specifically, according to the standard DPO formulation (Equation 3), the implicit reward of MLLMs
 534 can be expressed as follows:

$$\begin{cases} r(m, x, y) = \beta \log \frac{\pi_{\theta}(y|m, x)}{\pi_{ref}(y|m, x)} + \beta \log Z(m, x), \\ Z(m, x) = \sum_y \pi_{ref}(y|m, x) \exp\left(\frac{1}{\beta} r(m, x, y)\right), \end{cases} \quad (12)$$

535 where m , x , and y denote the input image, textual prompt, and corresponding response, respectively.
 536 π_{θ} , π_{ref} , and r are the policy model, reference model, and implicit reward function, respectively.
 537 $Z(m, x)$ is the partition function for the multimodal scenario, derived by incorporating an image
 538 variable into the single-modal partition function defined by standard DPO [9].

539 Adopting the mainstream vision-oriented preference learning paradigm that introduces additional
 540 contrastive images for preference alignment, we derive the following loss function by substituting the
 541 above implicit reward formulation into the Bradley-Terry model (Equation 1),

$$\begin{aligned} \mathcal{L}_{VCO}^* &= -\log \sigma(r(m_w, x, y_w) - r(m_l, x, y_w)) \\ &= -\log \sigma\left(\beta \log \frac{\pi_{\theta}(y_w|m_w, x)}{\pi_{ref}(y_w|m_w, x)} - \beta \log \frac{\pi_{\theta}(y_w|m_l, x)}{\pi_{ref}(y_w|m_l, x)} \right. \\ &\quad \left. + \beta \log Z(m_w, x) - \beta \log Z(m_l, x)\right). \end{aligned} \quad (13)$$

542 Existing methods directly cancel out $Z(m_w, x)$ and $Z(m_l, x)$ in Equation 6, which is apparently
 543 inappropriate based on the above rigorous derivation.

544 To better understand the role of partition functions in the preference learning process, we calculate
 545 the gradient of \mathcal{L}_{VCO}^* with respect to θ . To facilitate the gradient calculation, we first define:

$$\begin{cases} u = \beta \log \frac{\pi_{\theta}(y_w|m_w, x)}{\pi_{ref}(y_w|m_w, x)} - \beta \log \frac{\pi_{\theta}(y_w|m_l, x)}{\pi_{ref}(y_w|m_l, x)}, \\ c = \beta \log Z(m_w, x) - \beta \log Z(m_l, x). \end{cases}$$

546 where u involves the policy model optimization, while c does not. In fact, c remains constant across
 547 different policy model parameters θ because $Z(m, x)$ is independent of θ in its calculation.

548 Accordingly, the gradient of \mathcal{L}_{VCO}^* with respect to θ is then given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}_{VCO}^*}{\partial \theta} &= \frac{\partial \mathcal{L}_{VCO}^*}{\partial u} \cdot \frac{\partial u}{\partial \theta} \\
&= -\frac{\partial}{\partial u} \log \sigma(u + c) \cdot \frac{\partial u}{\partial \theta} \\
&= -(1 - \sigma(u + c)) \cdot \frac{\partial u}{\partial \theta} \\
&= -\sigma(-(u + c)) \cdot \frac{\partial u}{\partial \theta}.
\end{aligned} \tag{14}$$

549 Similarly, we can derive the gradient of the visual-oriented contrastive objective used in existing
550 works (i.e., Equation 6) as:

$$\frac{\partial \mathcal{L}_{VCO}}{\partial \theta} = -\sigma(-u) \frac{\partial u}{\partial \theta}. \tag{15}$$

551 By comparing the gradients in Equations 14 and 15, we observe that the constant c in Equations 14
552 acts as an offset modulating the coefficient of the gradient term $\frac{\partial u}{\partial \theta}$. Intuitively, since $Z(m, x)$
553 integrates over all possible responses y , it inherently captures the global quality of the response space
554 conditioned on (m, x) . Thus, c reflects the reference model’s global response quality discrepancy
555 between the contexts (m_w, x) and (m_l, x) . Specifically, a larger value of c indicates a stronger
556 discrepancy in model behavior between (m_w, x) and (m_l, x) , exerting greater influence on gradient
557 adjustment for preference alignment optimization. Conversely, a smaller value of c implies a weaker
558 response quality gap between the multimodal context pair, less impacting the gradient update during
559 training. However, existing methods neglect the two partition functions in visual-oriented contrastive
560 optimization. In essence, these methods assign the same value (i.e., $c = 0$) to all contrastive samples,
561 preventing adaptive weight adjustment for different image contexts. Consequently, the model fails
562 to achieve optimal visual understanding capability, as it tends to either over-attend to images with
563 limited informative signals or under-reason about images containing complex visual clues.

564 C Case Study

565 In this part, we present the case study on our constructed preference pair data for model training and
566 generated responses for model evaluation.

567 **On Preference Pair Construction.** As illustrated in Figure 1, each training sample for our symmetric
568 preference optimization comprises two triplets sharing the same prompt and similar but distinct image-
569 response pairs. Empirically, the contrastive images are identified using CLIP similarity scores, while
570 the responses are generated using the caption-anchored response preference pair construction pipeline
571 described in Section 4.1 for corresponding images given the same prompt. Figure 5 illustrates
572 two training examples from our constructed preference dataset, where claims in each response that
573 contradict (and are thus potentially hallucinated relative to) the paired response are highlighted in red.
574 These two examples demonstrate that the effectiveness of using CLIP similarity to obtain contrastive
575 image pairs with subtle visual differences, and confirm that the corresponding responses generated
576 by our preference data construction pipeline indeed exhibit strong linguistic alignment with only
577 minor claim variations. Overall, these examples show the effectiveness of our training preference
578 pair construction strategy.

579 **On Hallucination Mitigation.** To demonstrate the advantages of SymMPO, we compare the
580 specific responses generated by original LLaVA-1.5-7B and our SymMPO-enhanced version with
581 samples from two widely used benchmarks, MMHal-Bench and Object-HalBench. Both benchmarks
582 feature diverse visual questions closely related to daily life. Figure 6 and Figure 7 present the
583 corresponding results, with hallucinated content highlighted in red. As can be seen, LLaVA-1.5-
584 7B+SymMPO consistently generates more accurate responses than the original LLaVA-1.5-7B across
585 both benchmarks. Even for visual questions requiring detailed responses in Object-HalBench, LLaVA-
586 1.5-7B+SymMPO maintains precise, hallucination-free responses, while LLaVA-1.5-7B fails by
587 producing responses with hallucinations. These results confirm the effectiveness of our SymMPO in
588 enhancing vision-language understanding and mitigating MLLM hallucinations.

Case 1**Prompt:** How do these people know each other?**Original Image****Contrastive Image**

Response 1 (For Original Image): These people know each other as they are a family, **consisting of two men and a woman**, who have gathered around a wooden dining table to share a meal and spend quality time together. They are sitting on three different chairs (green, yellow-green, and red), and **the table is set with various items such as plates, bowls, knives, spoons, glasses, and condiments/bottles**. This indicates that they are enjoying a dinner together, which is a common activity for families to bond and strengthen their relationships.

Response 2 (For Contrastive Image): The group of people in the image, **consisting of two men, a woman, and a child**, are likely related or close friends, as they are gathered together for a meal. They are sitting **around a dining table with two dark-colored bottles**. The presence of food and drinks, coupled with the comfortable seating arrangement, indicates that they are part of a social gathering or celebration. It is likely that they are enjoying each other's company and sharing a meal, which is a common way to bond and strengthen relationships.

Case 2**Prompt:** What are the main objects or subjects in the image?**Original Image****Contrastive Image**

Response 1 (For Original Image): The main objects or subjects in the image **are a dense forest with tall, leafy trees featuring vibrant green foliage, suggesting spring or summer**. There is a rocky outcropping composed of layered ledges in light gray to tan hues, with distinct horizontal stratification and partial moss coverage. **The ground is covered with grass, moss, and low-lying plants**, along with patches of darker soil near the base of the rocks.

Response 2 (For Contrastive Image): The main objects or subjects in the image **are trees, fallen leaves, and a decomposed log**. The scene features a forest with mostly slender to medium-sized trees, some with light-colored bark (possibly birch) and others with darker bark. **The ground is covered with a thick layer of brown and yellow leaves, indicating autumn, while some green foliage remains on the trees**. Sparse shrubs and small plants are visible near the base of some trees. **A fallen log lies horizontally in the mid-ground, partially decomposed and blending into the natural environment**. Soft, diffused sunlight filters through the canopy, creating dappled shadows on the ground, enhancing the picturesque autumn setting.

Figure 5: Two training samples from our built preference dataset used to optimize SymMPO, with hallucinated elements relative to the other response highlighted in red.

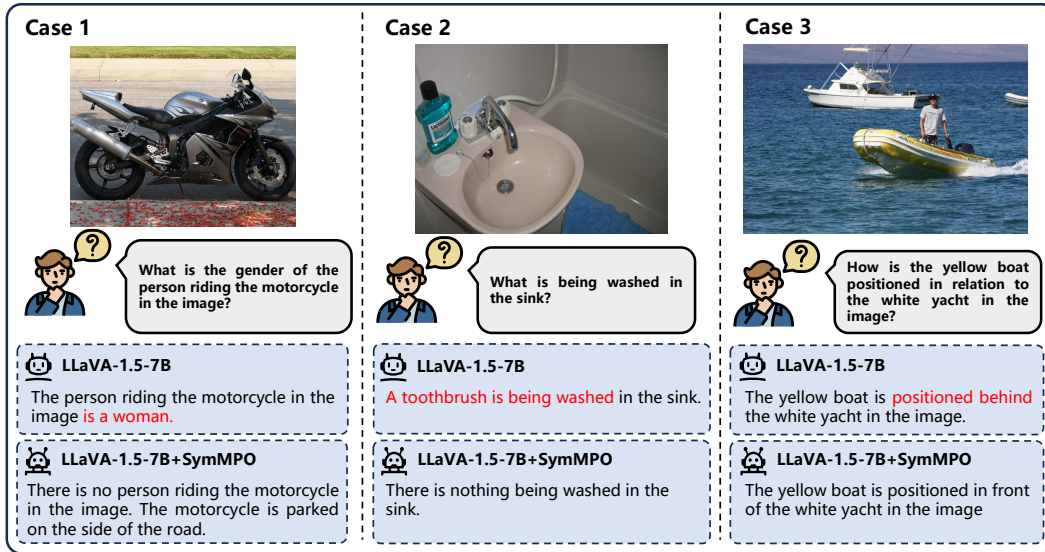


Figure 6: Responses generated by LLaVA-1.5-7B and our SymMPO-enhanced version for data examples from MMHal-Bench, with hallucinated contents highlighted in red.



Figure 7: Responses generated by LLaVA-1.5-7B and our SymMPO-enhanced version for data examples from Object-HalBench, with hallucinated contents highlighted in red.