

SMPRO: Self-Supervised Visual Preference Alignment via Differentiable Multi-Preference Multi-Group Ranking

Sirnam Swetha¹, Rui Meng², Shwetha Ram², Tal Neiman², Son Tran², Mubarak Shah^{1,2}

¹Center for Research in Computer Vision, University of Central Florida

²Amazon

swetha.sirnam@ucf.edu, {rmmeng, shweram, taneiman, sontran}@amazon.com, shah@crcv.ucf.edu

Abstract

Direct Preference Optimization (DPO) has emerged as a simple and effective approach for aligning models with human preferences. However, existing DPO-based methods suffer from 3 key drawbacks: they rely on only a single positive-negative preference pair per question, restricting the diversity and richness of feedback; they often emphasize minimizing negative preference scores while neglecting to strengthen the positive preferences; and they depend on either human-annotated preferences or expert model outputs - both expensive and difficult to scale. Moreover, the *deterministic* ranking assumptions of recent Group-based preference optimization methods break down in open-ended tasks such as Visual Question Answering (VQA), where multiple answers can be equally plausible but differ subtly in relevance or specificity. Given this subtle variance in preferences, we propose to perform ranking over groups of preferences rather than relying on fine-grained ranking of individual ones, which is often noisy and subjective. To address these challenges, we introduce Self-Supervised Visual Preference Alignment via Differentiable Multi-Preference Multi-Group Ranking (SMPRO), a novel framework that (1) self-generates rich, diverse preference groups while eliminating the need for external annotations, (2) employs a fully differentiable ranking objective based on sorting networks to capture nuanced preference gradients across arbitrary numbers of preferences both within and across these groups, and (3) incorporates multiple positive preferences to enrich the positive preference group, capturing subtle distinctions among high-quality preferences. Extensive experiments across diverse visual tasks show that our approach achieves state-of-the-art performance in self-supervised setting. Specifically, our model surpasses existing baselines, achieving notable gains such as 82.4% on MM-Bench, 63.2% on MMStar, 94.6% on LLaVA-W, and 81.9% on AI2D. These results underscore the effectiveness of our approach in capturing richer preference signals and demonstrate its scalability for open-ended, ambiguous VQA tasks.

Introduction

Recent advancements in large language models (LLMs) have led to the emergence of Vision Large Language Models (VLLMs)(Alayrac et al. 2022; Li et al. 2023a; Dai et al. 2023; Sirnam et al. 2024; Liu et al. 2024; Li et al. 2024; Kim et al. 2025), which are typically pretrained with abundant

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

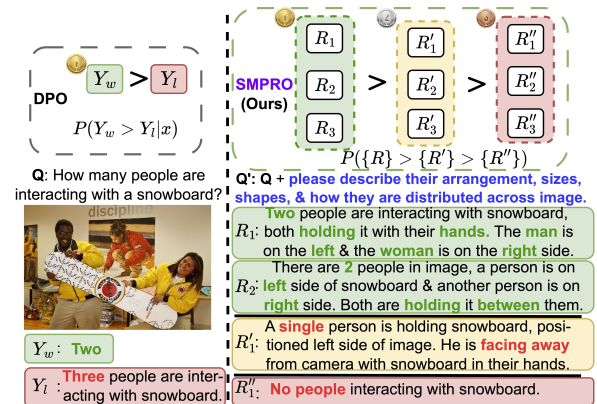


Figure 1: DPO (left) compares just one positive (Y_w) & negative (Y_l) response. In contrast, our SMPRO method (right) generates multiple positive and negative responses, groups them into ordered sets $\{R\} > \{R'\} > \{R''\}$, and optimizes group-level preferences. This enables a richer/more nuanced preference data in self-manner without any external models/human annotators. Responses from external model (Zhou et al. 2024) (left), right (ours).

image-text pairs for modality alignment and further fine-tuned with multimodal instruction data to equip the model with the ability to follow instructions accurately. However, instruction-finetuned models are not proficient at providing factually grounded content associated with images and often suffer from hallucinations (Zhao et al. 2023; Zhou et al. 2024). A potential reason for this limitation is their tendency to prioritize common sense or stereotypes present in the training language data. Recent research suggests that the problem arises from a lack of human feedback and emphasizes enhancing alignment through preference tuning, to reduce hallucinations and improve visual instruction capabilities (Sun et al. 2023; Zhao et al. 2023; Yu et al. 2024a; Zhou et al. 2024; Zhu et al. 2024).

RLHF (Christiano et al. 2017) and DPO (Rafailov et al. 2023) have emerged as prominent techniques in preference tuning. RLHF involves a complex, two-staged approach that requires training a reward model to score/rank responses and

then use this reward model to optimize the policy. (Sun et al. 2023) employs RLHF to enhance multimodal alignment. In contrast, DPO eliminates the explicit reward model requirement, simplifying and reducing computational costs. Recent studies (Zhou et al. 2024; Zhu et al. 2024) have proposed visual preference alignment approaches with DPO that involves (i) preference data generation, and (ii) preference-optimization. While POVID (Zhou et al. 2024) leverages expert and self model for generating two types of negative preferences, and perform two-stage preference optimization. SeVA (Zhu et al. 2024) proposes a simple self-supervised approach that leverages augmentations to generate negative preferences and perform one-stage preference alignment with pair of preferences. However, these methods primarily focus on negative preferences, overlooking positives and are limited to pairwise comparisons.

Group Relative Policy Optimization (GRPO) (Shao et al. 2024) offers an efficient reinforcement learning approach by estimating baselines from group scores without relying on explicit critic models, significantly reducing training complexity. Similarly, Vision-R1 (Huang et al. 2025) employs GRPO to enhance reasoning in multimodal large language models (MLLMs) through Reinforcement Learning (RL), relying heavily on deterministic scenarios such as multimodal math tasks. However, these deterministic methods face limitations in general Visual Question Answering (VQA) settings, where questions like “Describe the image” or “What is happening in this image?” typically have multiple plausible responses, with subjective and nuanced variations in relevance and specificity.

To address these limitations, we propose Self-Supervised Visual Preference Alignment via Differentiable Multi-Preference Multi-Group Ranking (SMPRO), that extends pair-wise contrasts to multi-group and multi-preference contrasts. However, in order to achieve this, 2 key challenges remain: (i) the inherent difficulty of generating diverse multiple preferences, and (ii) the consistent ranking of multiple preferences amidst inherent biases in human or model-based judgments (Wang et al. 2023; Stureborg, Alikaniotis, and Suhara 2024; Shi, Ma, and Vosoughi 2024). To tackle the issue of diverse response generation, we propose to incorporate reasoning explanations alongside model responses. This approach serves a dual purpose: it enhances response diversity while also leveraging explanation-based knowledge to enhance the model’s alignment, ultimately leading to more meaningful and well-structured responses. Ranking multiple preferences consistently is a challenging task. To alleviate this issue, we propose group-ranking instead of score-based preference ranking. For this, we perform multiple augmentations of image, and for each augmented image, we apply high temperatures alongside seeking reasoning explanations to generate multiple diverse responses for a given image-question pair. These responses are categorized into a single preference group, where rankings are determined based on the predefined noise threshold used for augmentation.

Additionally, we introduce a flexible preference loss that leverages differentiable sorting networks (Petersen et al. 2021) to rank preferences across multiple groups. Within the multi-group preference optimization framework, SMPRO

can effectively model both multiple negative & positive responses. This capability enables SMPRO to benefit from learning preferences through multiple contrasts (Fig.1), similar to insights from contrastive learning (Chen et al. 2020; Radford et al. 2021; Swetha et al. 2021, 2023), where large-batch size contributes to improved representation quality (Song et al. 2024). Additionally, prior studies (Khosla et al. 2020; Han, Xie, and Zisserman 2020; Dwibedi et al. 2021; Kamath et al. 2024) have shown that better representation learning can be achieved by incorporating more positives. SMPRO naturally integrates multiple high-quality responses into its high-ranked groups.

We summarize the main contributions as:

- We propose SMPRO, the *first* self-supervised multi-group preference ranking framework that simultaneously incorporates multiple positive and negative preferences.
- Our framework eliminates reliance on human annotators/expert models by autonomously generating structured preference groups via noise-based augmentations.
- We propose a *flexible* differentiable sorting-based loss, facilitating preference optimization across multiple groups, regardless of their sizes.
- Our *flexible* ranking mechanism mitigates ambiguity among similar responses within a group, removing the need for explicit ranking of individual preferences.

Our method outperforms the state-of-the-art self-supervised approach (Zhu et al. 2024) by 27.4 margin and the base model (Li et al. 2024) by 34 on the MME Cognition benchmark. Extensive evaluations across various benchmarks further confirm the effectiveness of our approach.

Related Works

Preference Alignment. Recently, RLHF (Bai et al. 2022; Ouyang et al. 2022; Ziegler et al. 2020; Wu et al. 2021; Menick et al. 2022) and Direct Preference Optimization (Rafailov et al. 2023; Zhou et al. 2024) have emerged as effective alignment techniques. DPO simplifies RLHF by analytically expressing the reward function $r(x, y)$ in terms of the policy $\pi_\theta(y|x)$ and a reference model $\pi_{\text{ref}}(y|x)$, enabling directly optimize the policy model on preference data without training a reward model. In the DPO setting, the reward model is represented as $r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$, where β is constant and $Z(x)$ is partition function. The policy model is directly optimized on the positive (y_w) and negative (y_l) preference by minimizing the loss.

$$L = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right] \quad (1)$$

However, DPO (Rafailov et al. 2023) primarily emphasizes on pair-wise preference optimization. It largely overlook the utility of multiple positive preferences, an area shown to improve feature alignment in other machine learning domains. The idea of using both positive and negative preferences aligns with the concept of balanced feedback in

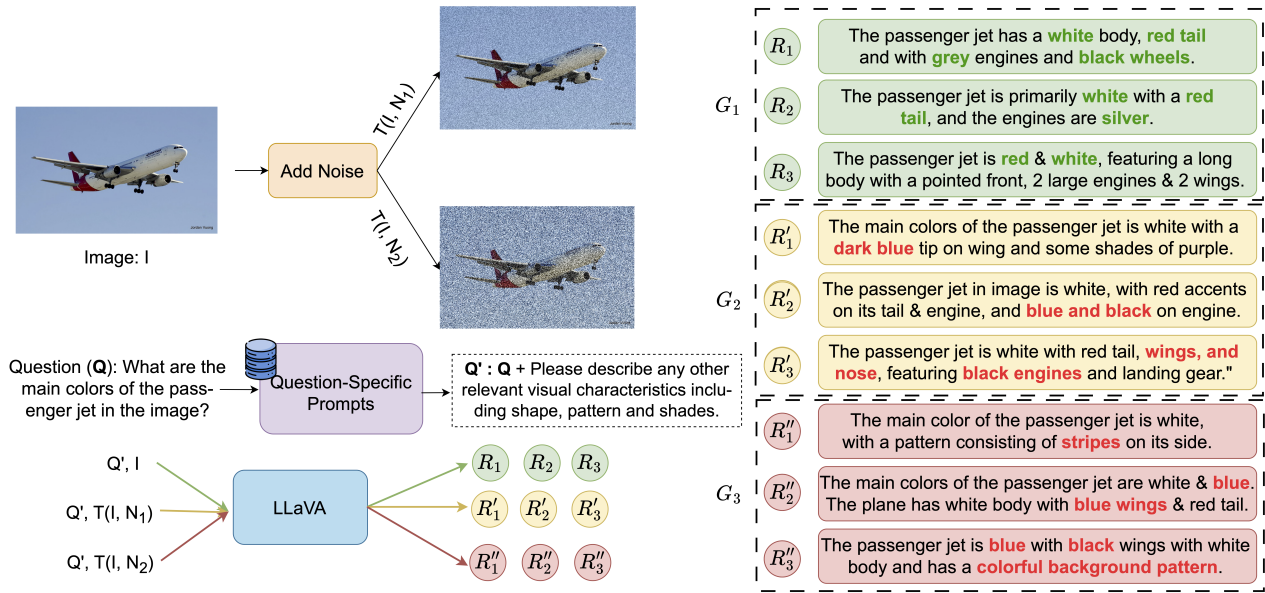


Figure 2: Self-Supervised Multi-Preference Data Generation Pipeline. Given an image (I) and question (Q), a question-specific prompt Q' is appended to encourage diverse responses by guiding the model to generate richer explanations by focusing on relevant aspects. The image is then subjected to controlled augmentations by applying different levels of noise (N_1, N_2). For each image and question pair, we employ the base model to generate multiple responses, capturing different visual aspects. These responses are grouped into structured preference groups G_i , according to augmentation intensity ($N_2 > N_1$). We perform ranking at the group level $G_1 > G_2 > G_3$, rather than on individual responses.

contrastive learning, enhancing feature robustness and reducing model biases. Contrastive learning techniques like SwAV (Caron et al. 2021) and multi-preference optimization (Zhou et al. 2023; Liao, Chu, and Wang 2024) have set the groundwork for expanded preference handling by promoting diverse preference representation and grouping techniques. Despite this, flexible preference grouping and ordering in visual reasoning remain under-explored. Our approach addresses this gap by not only expanding the preference set beyond binary distinctions but also implementing self-supervised grouping and ordering, without relying on pre-trained models or costly human annotations.

On the other hand, collecting preference from human annotation is expensive. Recent works collect the preferences from AI feedback. That significantly makes the preference data generation scalable. However, Using the LLM as a judge introduces the bias from the pretrained model (Stureborg, Alikaniotis, and Suhara 2024; Wang et al. 2023; Shi, Ma, and Vosoughi 2024). Those AI feedback methods would target specific domain to collect preference data (Zhao et al. 2023; Sun et al. 2023; Zhou et al. 2024) which may not transfer directly to other domains. To address those drawbacks, (Zhu et al. 2024) propose the first self-supervised framework for pair-wise preference alignment. However, they do not generate preference set. We propose flexible self-supervised approach to generate multi-group preference data without resorting to any AI model/human feedback.

Recent studies has demonstrated that prompting LLMs to engage in a step-by-step reasoning process can yield bet-

ter results than directly answering questions in a single step (Wei et al. 2023; Kojima et al. 2023; Zhong et al. 2024). This approach of “chain-of-thought prompting” encourages the model to work through complex problems systematically, improving accuracy and response quality. Our proposed SMPRO leverages the reasoning process to not only increase the diversity of responses but also incorporate the reasoning knowledge in preference data, which enhances the model reasoning in preference learning.

Recent study (Hejna et al. 2023) shows that DPO is a special case of contrast preference learning. Also, current DPO methods focus on generating good negative responses, they disregard other positive responses. (Kamath et al. 2024) show the importance of positives in contrastive learning and more specifically when performing contrastive negative-finetuning. Our SMPRO allows us model both positives and negatives, enhancing the efficient preference learning.

Differentiable Sorting and Ranking. Differentiable sorting techniques enable end-to-end training of neural networks using ranking or ordering supervision (Grover et al. 2019; Prillo and Eisenschlos 2020; Petersen et al. 2021, 2022). The focus is on having a differentiable relaxation for the sorting operator, which can be seen as a function that returns a permutation matrix which represents the permutation necessary to sort the given input sequence *i.e.* the sorted output is obtained by multiplying the permutation matrix with the input sequence. Most recently, differentiable sorting has been applied in self-supervised representation learning as an alternative for contrastive learning, with group-wise ordering

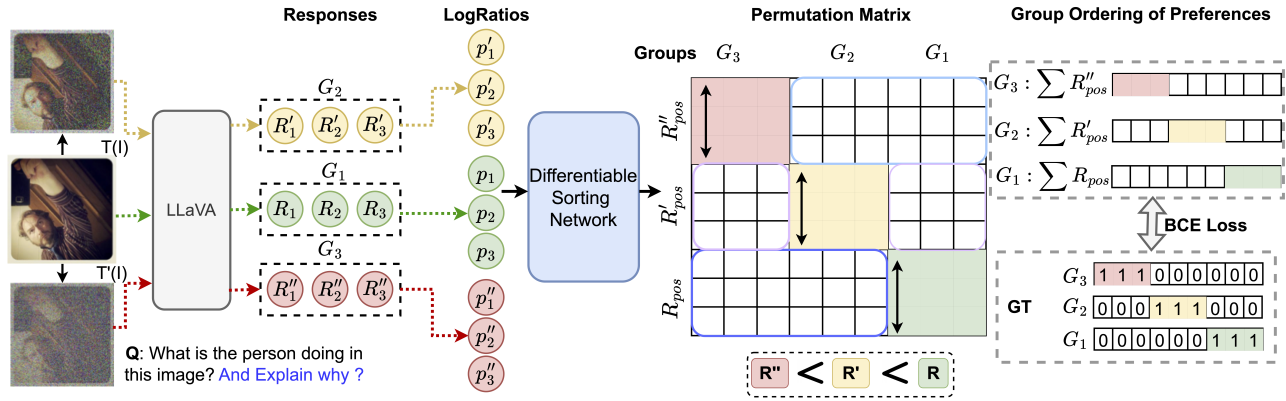


Figure 3: SMPRO Overview. Given an image (I) and question (Q), we create two augmented variants $T(\cdot)$ and $T'(\cdot)$, with T' stronger than T . We then append a reasoning prompt to Q and get diverse responses from the base model for images $I, T(I), T'(I)$, resulting in groups G_1, G_2, G_3 respectively. Then we compute the log ratios for each responses by subtracting the log probabilities of base (reference) model from those of the policy model. Next, we apply the differentiable sorting network on those log ratios to get a permutation matrix in which the column values denote the probabilities of sorted elements. Given, the preference ranking $R > R' > R''$ we sum over the groups G_1, G_2, G_3 rows of the permutation matrix shown as $\sum R_{pos}, \sum R'_{pos}, \sum R''_{pos}$ respectively. Finally, we calculate the multi-group preference (MGP) loss (see Eq. 3), that is the sum up of the BCE losses between group sum and the ground truth for three groups.

objectives (Shvetsova et al. 2023). To the best of our knowledge, our proposed SMPRO is the first work to leverage differentiable sorting for preference alignment.

Approach

Our proposed SMPRO approach consists of two key components: (1) We introduce a novel pipeline for generating multi-group preferences. This generation is fully self-supervised and scalable with no need for human involvement or AI feedback. Further, by introducing the explanations for the responses, the generated responses are meaningful and diverse, that contributes to better preference alignment. (2) We propose a flexible multi-group preference optimization framework by leveraging the differential sorting network (Petersen et al. 2021) and Kullback Leibler (KL) regularization. This framework allows us to model preferences across multiple groups with varying group sizes. First, we discuss the preference data generation pipeline below, followed by our proposed multi-group optimization framework.

Multi-Group Preference Data Generation

We propose a self-supervised preference generation approach capable of producing both multiple positive and negative preferences. To mitigate ambiguities among similar preferences, we group responses and assign a shared ranking within each group. This approach reduces intra-group variability while leveraging preference ranking across groups to enrich preference learning. Fig. 2 illustrates the significance of multiple preferences, specially in self-supervised setting. The motivation for generating multiple preference within the positive group comes from the enrichment of diverse preference contrasts. Specifically, since the base model is not

as robust as a strong AI model or human, its response to a question may not capture all the details in a single response as shown in Fig. 1 (left). Instead of pursuing the quality of strong positive responses, we collect multiple weak positive responses to cover the comprehensive details as shown in Fig. 2. These weak positive responses would lead to multiple diverse preference contrasts between the positive group and other negative group in preference optimization, leading to better and more robust model performance. The same motivation applies to negative preference generation. We first apply image augmentations, such as adding diffusion gaussian noise to the image (Zhu et al. 2024; Zhou et al. 2024), and then generate responses using model itself with high temperature to get negatives.

Generating Meaningful Diverse Responses Generating diverse responses is challenging, in particular generating those responses from the base model itself using high temperature sampling (Zhu et al. 2024; Zhang et al. 2024). To address this issue, we introduce the reasoning process into the prompt, asking model to provide the thoughts of reasoning. It allows us to not only enrich the diversity of responses but also incorporating reason knowledge in the preference to boosting the model reasoning capability in the preference alignment. Improving the model reasoning capability is necessary since most models may answer the question correctly but the underlying reasoning is not consistent when responding to questions in VQA tasks (Selvaraju et al. 2020). In summary, incorporating the reasoning process serves for two key purposes: (i) it integrates reasoning into preference optimization, which helps enhance the model’s reasoning consistency, and (ii) it enables the generation of diverse responses across all question types.

Generating Preference Ranking By leveraging the reason-

ing process, our preference data generation is scalable for multiple diverse responses. But ranking those responses consistently is difficult. Current works (Li et al. 2023b), leverage LLM to score the responses with different principles in different dimensions, and then get an average score for ranking. These overall scores can not generate consistent ranks across responses. In particular, for similar responses, the preference rankings are highly sensitive to the prompt. Even with the same prompt, the preferences may be annotated differently across runs. This inconsistent ranking behavior has been studied in (Shi, Ma, and Vosoughi 2024; Stureborg, Alikaniotis, and Suhara 2024; Wang et al. 2023). Those works leverage AI models as judge and find the preference rankings from AI models are not always fair, inconsistent and has biases.

To address the inconsistency of preference ranking, we propose a self-supervised approach that does not rely on AI models and human annotators. Moreover, our proposed approach can handle the trade-off between the learning from preference ranking and alleviating the ambiguity issue among the responses. Our approach is described below.

Given image I and question Q , we generate preference with three groups denoted by $G = \{G_1, G_2, G_3\}$. Here, G_1, G_2, G_3 denote responses generated with original image I , augmented image $T(I)$ and stronger augmented image $T'(I)$ respectively as shown in Fig. 3. Within each group, there are multiple responses generated from the same prompt but different augmented versions of the image using model itself. For example, we collect three samples for each groups denoted by $G_1 = \{R_1, R_2, R_3\}$, $G_2 = \{R'_1, R'_2, R'_3\}$ and $G_3 = \{R''_1, R''_2, R''_3\}$. Moreover, to reduce the ambiguity of preference ranking, we assume the preference within each group is the same and preference across groups following $R_i > R'_j > R''_k$ for any index i, j, k in the corresponding group. Then in the next step, we introduced a novel multi-group preference framework. We show qualitative examples of the preference dataset applying our proposed multi-group preference data generation highlighting the benefits of multiple preferences for both positive and negative responses in Fig. 5.

Multi-Group Preference Optimization

We leverage differentiable sorting networks (Petersen et al. 2021) and Kullback-Leibler divergence regularization to propose a new framework, that allow us conduct preference optimization for multi groups with varying sizes. To introduce our proposed multi-group preference optimization, first we cover preliminaries about differentiable sorting networks and then present our method.

Differentiable Sorting Networks The differential sorting networks (DSN) belongs to a category of sorting algorithms found in traditional literature (Knuth 1998), rather than a neural network. In this work, we utilize the odd-even sorting network, an algorithm that sorts by comparing and swapping neighboring elements at alternating odd and even indices. Given an input sequence of length L , the algorithm sorts them in ascending order in L steps. Specifically, the sorting network is defined as a composition of

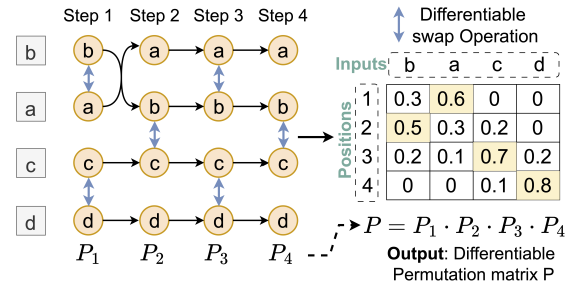


Figure 4: Overview of differentiable sorting network with odd-even sorting. At each step, it alternately compares neighboring elements at odd/even indices and applies a differentiable swap operation if the elements are out of order. The swap operation for each step s defines a differentiable permutation matrix P_s , in which the column values denote the probability of sorting the elements to the corresponding position/rank. The final permutation matrix, is defined as the multiplication of matrices obtained at each step s .

functions, with each function representing a sorting step in which pairs of elements in the input sequence are compared and swapped according to a differentiable swap operation as shown in Fig. 4. The differentiable swap operation relaxes the conditional swap operation. For elements (d_i, d_j) where $i < j$, the discrete swap is defined as $d'_i = \min(d_i, d_j)$ and $d'_j = \max(d_i, d_j)$. The differentiable relaxation of this, as proposed in (Petersen et al. 2021), approximates this swap operation in a way that allows for gradient flow, enabling optimization through backpropagation as shown in Eq. 2.

$$\begin{aligned} d'_i &= \text{softmin}(d_i, d_j) = d_i f(d_j - d_i) + d_j f(d_i - d_j), \\ d'_j &= \text{softmax}(d_i, d_j) = d_i f(d_i - d_j) + d_j f(d_j - d_i) \end{aligned} \quad (2)$$

where $f(x) = \frac{1}{\pi} \arctan(\beta x) + 0.5$, and $\beta > 0$ is a hyper-parameter. When $\beta \rightarrow \infty$ the differential swap operation converges to discrete swap operation. Then, the differentiable *conditional* swap operation for the elements (d_i, d_j) is defined as a permutation matrix $P_{\text{swap}(d_i, d_j)} \in \mathbb{R}^{L \times L}$, which is an identity matrix except for entries $P_{ii}, P_{ij}, P_{ji}, P_{jj}$ where $P_{ii} = P_{jj} = f(d_j - d_i)$ and $P_{ij} = P_{ji} = f(d_i - d_j)$. The permutation matrix P_s for step s is the product of permutation matrices corresponding to independent swap operations, $P_s = \prod_{i \in R_\Theta} P_{\text{swap}(d_i, d_{i+1})}$, where Θ is the set of odd indices if s is odd and the set of even indices if s is even. The final permutation matrix P is defined as product of all permutation matrices *i.e.* $P = P_1 \cdot P_2 \cdot \dots \cdot P_L$. In the relaxed version, the column values can be seen as distribution over possible positions for element, which in discrete case, is exactly one entry indicating position corresponding to the which column it is placed. Given the correct order of input elements, the sorting loss is defined as $L_{\text{sort}} = \frac{1}{L} \sum_{i,j} \text{BCE}(P_{ij}, Q_{ij})$, where L is the number of inputs; Q is the ground truth permutation matrix and BCE represents binary cross-entropy loss function.

Method	MME	MMBench	MMStar	LLaVA-W	AI2D
Expert Models					
GPT-4V	517	75	-	57.1	98.0
GPT-4o	-	-	-	-	106.1
Baselines					
LLaVA-1.5-7B(L) (Liu et al. 2024)	348.2	64.3	58.3	33.3	63.4
LLaVA-OV-Qwen2-0.5B(OV) (Li et al. 2024)	240	52.1	-	37.5	74.2
LLaVA-OV-Qwen2-7B(OV) (Li et al. 2024)	418	80.8	-	61.7	90.7
Expert/Human Feedback Approaches					
L + RLHF (Sun et al. 2023)	360.2	60.4	-	33.0	63.7
L+ RLAIIF-V (Yu et al. 2024b)	302.9	62.6	-	35.4	72.7
POVID (Zhou et al. 2024)	350.4	64.9	54.9	-	68.7
Self-Supervised Approaches					
SMPRO (Ours)-0.5B	254	53.2	47.2	37.4	75.1
SeVa-7B (Zhu et al. 2024)	305	64.5	59.2	-	72.2
OV-7B + SeVa	424.6	81.1	80.5	62.2	90.4
SMPRO (Ours)-7B	452	82.4	80.6	63.2	94.6

Table 1: Comparison with SoTA using same base model on various benchmarks. OV: LLaVA-OneVision, L: LLaVA-1.5

MGPO Framework The overall approach is illustrated in Fig. 3. As shown, we first generate three multi-group preference denoted as $[R]_i, [R]'_i, [R]''_i$ and compute the corresponding log ratios, the ratio of logprobs between the reference and alignment model *i.e.* $\log \frac{\pi_\theta(R|x)}{\pi_{ref}(R|x)}$. Then we applied differential sorting network on those log ratios to obtain the probabilistic position vector for groups. Finally, we compute the BCE loss between generated position vector and ground truth for each group, and sum them up for our final loss. Specifically, assuming we have n preference groups with varying size g_1, g_2, \dots, g_n and the generated permutation matrix is P , we propose the multi-group preference loss (MGP) is displayed as Eq. 3.

$$L_{MGP} = \frac{1}{N} \sum_{i=1}^N (BCE(\sum_{k=c_{i-1}}^{c_i} P_{k,i}, \mathbb{1}_{c_{i-1} \leq i \leq c_i})) \quad (3)$$

where $N = \sum_{i=1}^n g_i$ is total number of preferences, $c_0 = 1$ & $c_k = \sum_{i=1}^k g_i$ for $k > 0$ are cumulated index of group preferences. $P_{k,i}$ represents the probability for assigning the element i in list to position k and $\mathbb{1}$ is an indicator function.

Note that when we consider two groups, where both the positive and negative group contain only a single preference, our MGP loss becomes *equivalent* to the DPO loss. The training loss trajectories match closely, showing nearly identical optimization behavior and convergence. Thus, when constrained to pairwise preferences, our sorting-based MGP loss reduces precisely to the DPO formulation. Similar to most of preference alignment methods (Xu et al. 2024), our proposed MGP loss also suffers from the over-optimization issue. To address this limitation, we propose to add a KL divergence constraint for the positive responses to not deviate from the reference model. Our proposed multi-group preference loss can be expressed in Eq. 4

$$L = L_{MGP} + \alpha * \frac{1}{|G_{pos}|} \sum_{i \in G_{pos}} \log \frac{\pi_\theta(r_i | x)}{\pi_{ref}(r_i | x)} \quad (4)$$

where G_{pos} refers to the positive preference group.

Experiments

In this section, we present our experimental setup, including datasets, benchmarks, and results. Our goal is to validate the effectiveness of multi-group preference ranking under a self-supervised setting.

Dataset and Setup. We construct our training dataset from the POVID split (Zhou et al. 2024), derived from the LLaVA-Instruct-150K dataset (Liu et al. 2024), which includes tasks like logical QA, captioning, and VQA. To simulate diverse visual conditions, we apply two levels of Gaussian diffusion noise with steps of 500 and 800, producing weakly and strongly augmented images I_{n500}, I_{n800} in addition to the original image I . For each image-question pair, we generate up to three diverse responses, resulting in three preference groups per sample and a total of nine preferences.

Implementation. We employ LLaVA-OneVision (Li et al. 2024) as our base model, and apply a differentiable sorting-based loss with a hyperparameter α set to 0.17, for scaling the regularization constraint. We use the AdamW optimizer with an initial learning rate of $2e-6$, weight decay set to 0. The learning rate follows a cosine annealing schedule.

Results

Our approach is compared against the state-of-the-art methods in preference alignment for vision-language models, grouped into expert-tuned, human-feedback, and self-supervised categories. All methods are built on the LLaVA baseline. As shown in Tab. 1, our method (Ours-7B) achieves the best performance across all benchmarks. For example, on the MME benchmark, SMPRO reaches a score of 452, outperforming SeVa-7B (Zhu et al. 2024) by +27.4 and the base model (Li et al. 2024) by +34 points. Additionally, we observe consistent gains on MMBench, MMStar, LLaVA-Wild, and AI2D, highlighting the benefits of incorporating group-level (flexible) multi-preference alignment. Moreover, reasoning prompts increase MATTR-50 by



Q: What is the cat doing in the image?
Answer the question and explain why?

Responses

- "The cat is lying on a **suitcase**, possibly a **black carry-on**, placed on the **floor**."
- "The cat is laying down in a **suitcase** on the **floor**, looking at the **camera**. It appears that the cat is curious and might be attracted to the suitcase",
- "The cat is sitting or laying **inside** a **suitcase** in the image."
- "The cat in the image is laying down on **top** of an **open suitcase**. The cat finds the suitcase to be a **comfortable** and familiar surface to rest on",
- "In the image, the cat is **sleeping** or **resting** on top of a suitcase.",
- "The cat is resting or **sleeping** on **top** of a **black suitcase** in the image."
- "The cat in the image is lying down on the **car**.",
- "The cat in the image is **resting** or lying on a piece of **furniture**, possibly a **car seat** or a **chair**.",
- "The cat is lying down on a **car seat**, specifically **resting** in the back of the **vehicle**."

Figure 5: Qualitative sample from our preference dataset. Each dotted box shows a preference group, with responses capturing different aspects of the scene. Grouping multiple responses enhances the richness and diversity of preferences in each group, in contrast to using a single response. This is specifically crucial for self-supervised setting as the base model is not as strong as external models, which are utilized for generating high quality preference data. Including multiple weaker preferences within groups leads to more diverse contrast in preference learning, allowing the model better learn multiple nuances in the scene without reliance on external models.

Preferences			MME_C	LLaVA-W
G_1	G_2	G_3		
1	2	2	243	74.4
2	2	2	247	74.5
3	3	3	254	74.9

Table 2: Effect of Number of Preferences per Group.

Groups	Preferences			MME_C	LLaVA-W
	G_1	G_2	G_3		
2	3	6	0	238	74.2
3	1	2	2	243	74.4
3	3	3	3	254	74.9

Table 3: Effect of Number of Preference Groups.

19.3% over non-reasoning prompts, demonstrating quantitative preference diversity gains from reasoning alone.

Ablation Studies

We use the LLaVA-OneVision-0.5B model for ablations.

Impact of Preference Group Size. We analyze the effect of varying the number of preferences within each group. As shown in Tab. 2, increasing group size improves performance, with the best results achieved when using 3 preferences per group. This indicates that richer intra-group comparisons help the model capture more nuanced differences.

Impact of # Groups. We also evaluate the effect of using different numbers of preference groups. Results indicate that using three groups yields the best score, as shown in Tab. 3. Increasing the number of groups allows the model to capture a broader diversity of preferences, which contributes to a richer understanding and robust preference alignment.

Qualitative Analysis To better understand the behavior of

our model under multi-group preference ranking, we present further qualitative examples from our dataset in Fig. 5.

Each dotted box represents a preference group, with responses generated under different levels of noise added to the image. Green group responses indicate top-ranked preferences, yellow/red denote lower-quality responses. We observe clear trends in response quality as noise increases. As shown in the Fig. 5, the question asks about the activity of the cat. The top group (green) provides consistent, accurate answers like “cat is lying on a suitcase” correctly identifying objects (suitcase, floor) and actions (resting, lying). As we move to the middle group (orange), we begin to see subtle confusions, for example, the mention of “sleeping on a suitcase” which is visually plausible but incorrect. The lowest-ranked group (red) includes more hallucinated or ambiguous phrases like “lying on a car seat” or “in the back of the vehicle” which are inconsistent with the actual scene, reflecting degradation due to higher image noise.

Summary

We presented SMPRO, a self-supervised framework for multi-group visual preference alignment. Unlike prior DPO-based methods that rely on single pairwise comparisons, SMPRO models multiple positive and negative responses per sample and ranks them flexibly at the group level. Our method generates diverse preferences using noise-based augmentations and reasoning prompts, and optimizes them with a flexible differentiable sorting-based loss. The preference groups in SMPRO are fully flexible, both in number and size allowing the model to handle varying levels of response diversity. This eliminates the need for human or expert annotations, captures nuanced preference signals, and improves alignment in open-ended VQA tasks. SMPRO achieves state-of-the-art results across multiple benchmarks, demonstrating its effectiveness, and contribution to self-supervised preference learning.

References

- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv:2204.05862*.
- Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2021. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv:2006.09882*.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*.
- Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
- Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Dwivedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9588–9597.
- Grover, A.; Wang, E.; Zweig, A.; and Ermon, S. 2019. Stochastic Optimization of Sorting Networks via Continuous Relaxations. In *International Conference on Learning Representations*.
- Han, T.; Xie, W.; and Zisserman, A. 2020. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 33: 5679–5690.
- Hejna, J.; Rafailov, R.; Sikchi, H.; Finn, C.; Niekum, S.; Knox, W. B.; and Sadigh, D. 2023. Contrastive preference learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*.
- Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Kamath, A.; Hsieh, C.-Y.; Chang, K.-W.; and Krishna, R. 2024. The hard positive truth about vision-language compositionality. *arXiv preprint arXiv:2409.17958*.
- Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.
- Kim, Y.; Swetha, S.; Kagdi, F.; and Shah, M. 2025. Safe-LLaVA: A Privacy-Preserving Vision-Language Dataset and Benchmark for Biometric Safety. *arXiv preprint arXiv:2509.00192*.
- Knuth, D. E. 1998. *The art of computer programming, volume 3: (2nd ed.) sorting and searching*. USA: Addison Wesley Longman Publishing Co., Inc.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. *arXiv:2205.11916*.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv:2408.03326*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; and Kong, L. 2023b. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.
- Liao, W.; Chu, X.; and Wang, Y. 2024. TPO: Aligning Large Language Models with Multi-branch and Multi-step Preference Trees. *arXiv:2410.12854*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Menick, J.; Trebacz, M.; Mikulik, V.; Aslanides, J.; Song, F.; Chadwick, M.; Glaese, M.; Young, S.; Campbell-Gillingham, L.; Irving, G.; and McAleese, N. 2022. Teaching language models to support answers with verified quotes. *arXiv:2203.11147*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155*.
- Petersen, F.; Borgelt, C.; Kuehne, H.; and Deussen, O. 2021. Differentiable Sorting Networks for Scalable Sorting and Ranking Supervision. In *International Conference on Machine Learning (ICML)*.
- Petersen, F.; Borgelt, C.; Kuehne, H.; and Deussen, O. 2022. Monotonic Differentiable Sorting Networks. In *International Conference on Learning Representations*.
- Prillo, S.; and Eisenschlos, J. 2020. Softsort: A continuous relaxation for the argsort operator. In *International Conference on Machine Learning*. PMLR.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Selvaraju, R. R.; Tendulkar, P.; Parikh, D.; Horvitz, E.; Ribeiro, M. T.; Nushi, B.; and Kamar, E. 2020. SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Shi, L.; Ma, W.; and Vosoughi, S. 2024. Judging the Judges: A Systematic Investigation of Position Bias in Pairwise Comparative Assessments by LLMs. *arXiv preprint arXiv:2406.07791*.
- Shvetsova, N.; Petersen, F.; Kukleva, A.; Schiele, B.; and Kuehne, H. 2023. Learning by Sorting: Self-supervised Learning with Group Ordering Constraints. *ICCV*.
- Sirnam, S.; Yang, J.; Neiman, T.; Rizve, M. N.; Tran, S.; Yao, B.; Chilimbi, T.; and Shah, M. 2024. X-Former: Unifying Contrastive and Reconstruction Learning for MLLMs. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland.
- Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; and Wang, H. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Stureborg, R.; Alkaniotis, D.; and Suhara, Y. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.
- Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; Keutzer, K.; and Darrell, T. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. *arXiv:2309.14525*.
- Swetha, S.; Kuehne, H.; Rawat, Y. S.; and Shah, M. 2021. Unsupervised Discriminative Embedding For Sub-Action Learning in Complex Activities. In *2021 IEEE International Conference on Image Processing (ICIP)*, 2588–2592.
- Swetha, S.; Rizve, M. N.; Shvetsova, N.; Kuehne, H.; and Shah, M. 2023. Preserving Modality Structure Improves Multi-Modal Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 21993–22003.
- Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv:2201.11903*.
- Wu, J.; Ouyang, L.; Ziegler, D. M.; Stiennon, N.; Lowe, R.; Leike, J.; and Christiano, P. 2021. Recursively Summarizing Books with Human Feedback. *arXiv:2109.10862*.
- Xu, H.; Sharaf, A.; Chen, Y.; Tan, W.; Shen, L.; Van Durme, B.; Murray, K.; and Kim, Y. J. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.
- Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; and Chua, T.-S. 2024a. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; and Sun, M. 2024b. RLAI-F-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness. *arXiv preprint arXiv:2405.17220*.
- Zhang, R.; Gui, L.; Sun, Z.; Feng, Y.; Xu, K.; Zhang, Y.; Fu, D.; Li, C.; Hauptmann, A.; Bisk, Y.; et al. 2024. Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward. *arXiv preprint arXiv:2404.01258*.
- Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond Hallucinations: Enhancing LVLMS through Hallucination-Aware Direct Preference Optimization. *arXiv:2311.16839*.
- Zhong, T.; Liu, Z.; Pan, Y.; Zhang, Y.; Zhou, Y.; Liang, S.; Wu, Z.; Lyu, Y.; Shu, P.; Yu, X.; Cao, C.; Jiang, H.; Chen, H.; Li, Y.; Chen, J.; Hu, H.; Liu, Y.; Zhao, H.; Xu, S.; Dai, H.; Zhao, L.; Zhang, R.; Zhao, W.; Yang, Z.; Chen, J.; Wang, P.; Ruan, W.; Wang, H.; Zhao, H.; Zhang, J.; Ren, Y.; Qin, S.; Chen, T.; Li, J.; Zidan, A. H.; Jahin, A.; Chen, M.; Xia, S.; Holmes, J.; Zhuang, Y.; Wang, J.; Xu, B.; Xia, W.; Yu, J.; Tang, K.; Yang, Y.; Sun, B.; Yang, T.; Lu, G.; Wang, X.; Chai, L.; Li, H.; Lu, J.; Sun, L.; Zhang, X.; Ge, B.; Hu, X.; Zhang, L.; Zhou, H.; Zhang, L.; Zhang, S.; Liu, N.; Jiang, B.; Kong, L.; Xiang, Z.; Ren, Y.; Liu, J.; Jiang, X.; Bao, Y.; Zhang, W.; Li, X.; Li, G.; Liu, W.; Shen, D.; Sikora, A.; Zhai, X.; Zhu, D.; and Liu, T. 2024. Evaluation of OpenAI o1: Opportunities and Challenges of AGI. *arXiv:2409.18486*.
- Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024. Aligning Modalities in Vision Large Language Models via Preference Fine-tuning. *arXiv:2402.11411*.
- Zhou, Z.; Liu, J.; Yang, C.; Shao, J.; Liu, Y.; Yue, X.; Ouyang, W.; and Qiao, Y. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*.
- Zhu, K.; Zhao, L.; Ge, Z.; and Zhang, X. 2024. Self-Supervised Visual Preference Alignment. In *ACM Multimedia 2024*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593.