SymDPO: Boosting In-Context Learning of Large Multimodal Models with Symbol Demonstration Direct Preference Optimization

Hongrui Jia[1,†] Chaoya Jiang[1,†] Haiyang Xu[2,*] Wei Ye[1,*] Mengfan Dong[1]
Ming Yan[2] Ji Zhang[2] Fei Huang[2] Shikun Zhang[1]
[1] National Engineering Research Center for Software Engineering, Peking University
[2] Alibaba Group
{jiahongrui, jiangchaoya, wye, zhangsk}@pku.edu.cn,
{shuofeng.xhy, fei.huang}@alibaba-inc.com

Abstract

As language models continue to scale, Large Language Models (LLMs) have exhibited emerging capabilities in In-Context Learning (ICL), enabling them to solve language tasks by prefixing a few in-context demonstrations (ICDs) as context. Inspired by these advancements, researchers have extended these techniques to develop Large Multimodal Models (LMMs) with ICL capabilities. However, existing LMMs face a critical issue: they often fail to effectively leverage the visual context in multimodal demonstrations and instead simply follow textual patterns. This indicates that LMMs do not achieve effective alignment between multimodal demonstrations and model outputs. To address this problem, we propose Symbol Demonstration Direct Preference Optimization (SymDPO). Specifically, SymDPO aims to break the traditional paradigm of constructing multimodal demonstrations by using random symbols to replace text answers within instances. This forces the model to carefully understand the demonstration images and establish a relationship between the images and the symbols to answer questions correctly. We validate the effectiveness of this method on multiple benchmarks, demonstrating that with SymDPO, LMMs can more effectively understand the multimodal context within examples and utilize this knowledge to answer questions better.

1. Introduction

The rapid advancement of Large Language Models (LLMs) [3, 7, 10, 37, 43, 43] has brought remarkable improvements in their In-Context Learning (ICL) capabilities [15]. By leveraging In-Context Demonstrations (ICDs), a small set of exemplars provided as context, these models achieve impressive performance on various language tasks. This breakthrough in Natural Language Processing (NLP) has catalyzed research efforts to extend similar contextual learning capabilities to Large Multimodal Models (LMMs) [1, 4, 23, 26]. The ultimate goal is to enable LMMs to effectively learn from a limited number of image-text pairs for specific tasks without parameter updates, thereby achieving few-shot learning in the multimodal domain.

To enhance the ICL capabilities of LMMs, prior works [4, 6, 23, 24, 26] have explored two primary approaches: The first approach [4, 6, 29, 42] involves pretraining LMMs

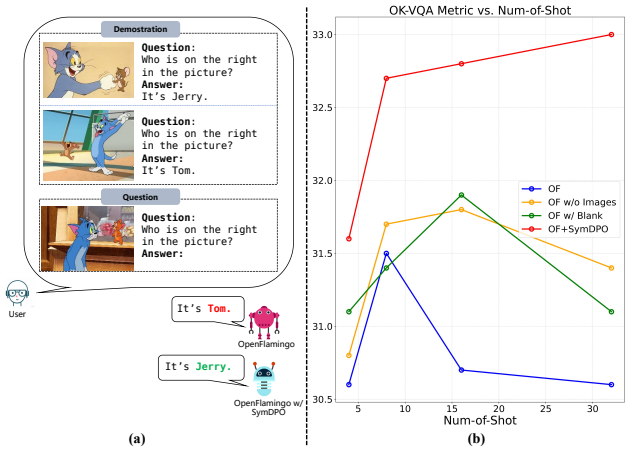

Figure 1. In subfigure (a), an example of visual context overlook is illustrated using OpenFlamingo as a case study. Here, OpenFlamingo [6] erroneously generates a response by solely following the textual cues in the demonstration, leading to an inaccurate answer. After applying SymDPO to enhance alignment, OpenFlamingo with SymDPO successfully corrects its response, accurately addressing the question. Subfigure (b) further demonstrates that for OpenFlamingo (OF), replacing images in the demonstration with blank placeholders (OF w/ blank) or omitting images altogether (OF w/o image) surprisingly yields even better performance than the original setup. This result suggests a substantial model dependency on textual context over visual information.

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

on massive-scale interleaved image-text data collected from the internet. The second approach [17, 21, 26] focuses on constructing specialized instruction-tuning datasets with numerous ICD examples. However, despite these efforts, recent studies [9, 11] have shown that both approaches still face a significant limitation, which we term **Visual Context Overlook.** This phenomenon manifests as the LMMs persistently struggle to effectively incorporate visual context from multimodal demonstrations, exhibiting a strong bias towards textual pattern matching. As illustrated in Figure 1 (a), when presented with multimodal examples, LMMs tend to generate responses by following textual patterns in the context while failing to properly utilize the critical visual information, leading to inaccurate responses. Additionally, as shown in Figure 1 (b), substituting images in ICDs with blank images or even removing them altogether does not affect model performance, further underscoring the limited role of visual information in the current alignment process of LMMs.

This issue highlights a core limitation in LMMs' ability to follow instructions from multimodal demonstrations within ICL scenarios accurately. Recently, Direct Preference Optimization (DPO) [39], a human preference-aligned reinforcement learning technique applied during the post-training phase, has been widely adopted to enhance LMMs' instruction-following capabilities [29, 32, 38, 45], offering a promising direction for addressing visual context overlook. However, current DPO methods exhibit two key limitations within ICL scenarios: **1. Insufficient Mechanisms for Multimodal In-Context Instruction Following:** Current DPO methods [38, 45] are largely optimized for general instruction-following tasks and lack the specialized mechanisms necessary to enhance LMMs' comprehension and adherence to the combined visual and textual information characteristic of multimodal demonstrations in ICL settings. **2. Challenges in Preference Data Collection for Visual Context Dependency:** In typical Visual Question Answering (VQA) tasks, many questions can be effectively answered based on text alone, without needing information from multimodal in-context demonstrations (ICDs). This reliance on textual cues creates a significant barrier to collecting reliable preference data for multimodal learning. Specifically, it complicates the distinction between "accepted" and "rejected" answers, as models may default to simple text pattern matching.

To overcome these limitations, we introduce **SymDPO** (Symbol Demonstration Direct Preference Optimization), a novel method specifically designed to compel LMMs to depend on both visual and textual inputs in ICDs by establishing a mapping between visual information and symbolic responses. SymDPO replaces traditional textual answers in ICDs with semantically neutral or mismatched symbolic text strings—specific characters or strings that have no semantic relevance to the visual context. This symbolic substitution compels the LMM to construct a mapping between visual elements and these symbolic strings, effectively linking the image content to a symbolic representation of the answer. As a result, the model can only generate correct responses by thoroughly interpreting the visual content within ICDs, as there is no relevant meaning in the symbolic text alone to support a response. This configuration makes visual information essential for understanding and responding accurately, ensuring that correct answers derive from a combined understanding of both image content and symbolic text. SymDPO thus redefines the model's reliance on visual context, reinforcing the multimodal comprehension required for accurate response generation in visually-dependent scenarios. Our contributions are as follows:

- We propose a novel symbolic preference optimization method, SymDPO, that compels LMMs to leverage multimodal information effectively, advancing their ability to integrate visual and textual cues within ICDs.
- We design and implement a pipeline that generates symbolic preference data, replacing textual answers with contextually mismatched symbols to enforce symbolic alignment with visual context.
- Through comprehensive experiments across multiple LMM architectures, we demonstrate consistent improvements in performance on various benchmarks, verifying SymDPO's effectiveness in addressing visual context overlook and enhancing multimodal understanding.

2. Related Work

2.1. In-Context Learning

In LLMs, prompt engineering handles specific tasks without constant fine-tuning. A variant of this method, ICL, enhances these abilities by generating prompts that incorporate multiple demonstrations [10, 15, 27, 28, 36]. ICL has already shown superior performance and broad applicability across many tasks [16, 18, 19, 34, 35] and can be easily modified for downstream tasks [2, 46, 48]. As LLMs continue to improve, more researchers are adapting them to the multimodal domain [30, 52, 54]. Leveraging the robust inference capabilities of LLMs, some LMMs have begun to display ICL capabilities, like Flamingo [4] and IDEFICS [23]. These models have notably improved their ICL abilities by incorporating multiple samples as contextual information during the training process. However, Multimodal ICL has some limitations: LMMs pay much more attention to the textual pattern instead of image information. Previous studies [49, 52] have primarily focused on the method of constructing context to mitigate the issue, overlooking the inherent characteristics of LMMs themselves, resulting in ineffective outcomes.

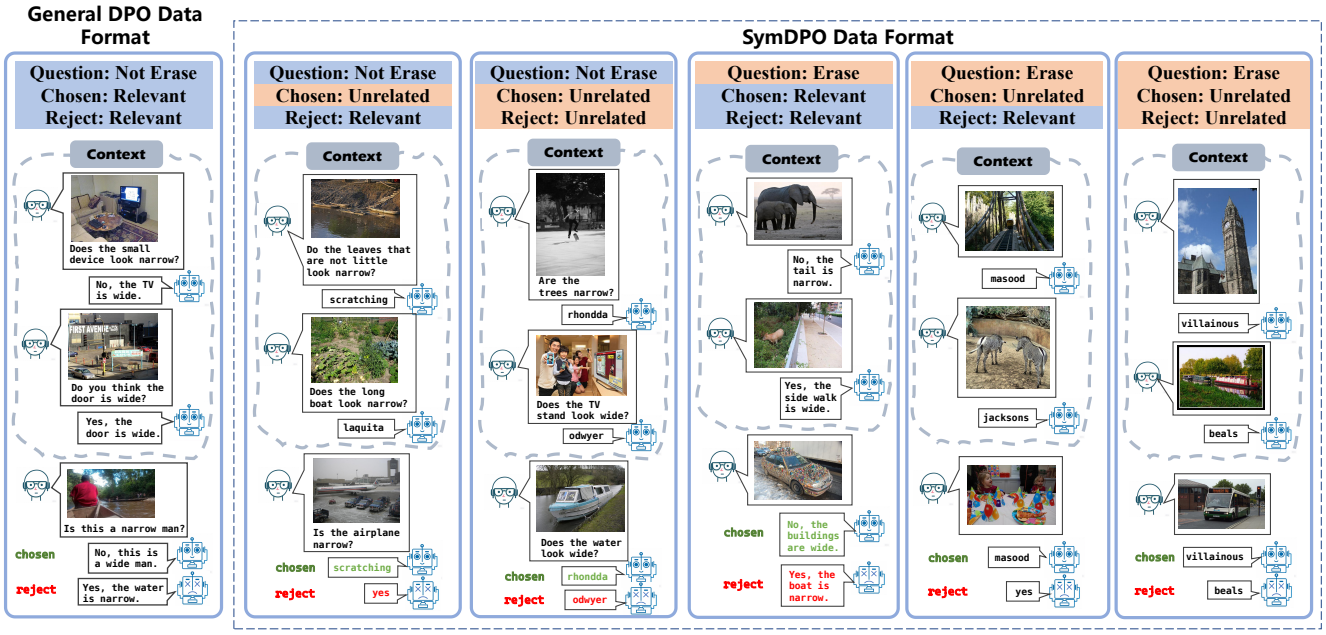

Figure 2. Comparison of General DPO and SymDPO Formats: General DPO relies solely on standard text for Questions, Answers, Chosen, and Rejected Answers, focusing on text-based training. In contrast, SymDPO replaces textual Answers with symbolized text to boost multimodal understanding, requiring models to interpret both visual and symbolized cues. This approach strengthens the model's ability to reason and decide in complex multimodal contexts.

2.2. Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) [37] has become a pivotal technique in guiding model responses to align more closely with human expectations. This approach fine-tunes models based on preference data obtained from human evaluators, facilitating the development of more intuitive and contextually accurate responses [22, 25]. Although RLHF has proven effective in aligning language models (LMs) with human values, its complexity and resource demands have prompted the search for alternatives. RAFT [14] chooses the best training samples through an existing reward model, while RRHF[51] uses a simpler ranking loss, maintaining the efficiency of PPO [40]. In contrast, DPO [39] directly optimizes LMs using a preference-based loss function, demonstrating improved training stability compared to traditional RLHF.

3. Method

To address the limitations of LMMs in leveraging image information within ICL, we propose SymDPO. As shown in Figure 2, SymDPO replaces the original answers in the context with unrelated words. This alteration prevents the model from depending solely on text patterns to infer answers. Instead, the model is driven to derive the correct answer by recognizing patterns that emerge jointly from the images and text in the context. In Section 3.1, we will discuss the construction of a dataset tailored for SymDPO to facilitate this balanced learning approach. Section 3.2 will elaborate on the specific steps and mechanisms of the SymDPO algorithm.

3.1. SymDPO Data Construction

The construction of the SymDPO data follows a three-step process. First, we gather and structure a QA dataset in an In-Context Learning (ICL) format. Based on this ICL dataset, we proceed to build a standard DPO dataset. Finally, we introduce the concept of SymDPO, expanding upon the general DPO data foundation to create the SymDPO dataset with enhanced multimodal challenges for LMMs.

Construct In-Context dataset: First, we collect image-question-answer triplets from VQA datasets such as GQA, VQAv2, and image classification datasets like ImageNet. These data points are then reorganized into an In-Context Learning (ICL) format. Questions that target similar task types, such as binary yes/no questions, or those related to image object categories, attributes, relationships, or quantities, are grouped together. For example, all questions in one group may revolve around the concept of "narrow" versus "wide." Next, for each category of questions, we construct data in the ICL format: $D_1, D_2, \ldots, D_N, F$, where $D_i = \{I_i, Q_i, A_i\}, i \in \{1, 2, \ldots, N\}$ and $F = \{\hat{I}, \hat{Q}, \hat{A}\}$. Here, D_i represents the i-th demonstration, and F denotes the final question-answer pair. We ensure that the demonstrations contain at least two different answers, with at least one answer matching the answer $\hat{A}$ in F. Building on this structure, we proceed to construct the DPO data required for training.

Construct Original DPO Dataset: For each constructed

ICL dataset, we treat the original answer $\hat{A}$ from the final QA round as the positive label. Then, we select a distinctly different answer A_j (where $A_j \neq \hat{A}$) from the same category of questions as the negative label, ensuring that this answer is not one of the previous answers. For instance, in the case of color-related questions, we may randomly choose another distinct answer like "The logo is black" as the negative label. With this, a single DPO data point is constructed.

Construct SymDPO Dataset: The SymDPO dataset is designed to increase the difficulty of answering questions, requiring Large Multimodal Models (LMMs) to fully comprehend the combined visual and textual information within the In-Context Demonstrations (ICDs) to accurately respond. Illustrated in the Figure 2, we constructed five distinct data configurations based on the standard DPO dataset to further enhance the model's comprehension capabilities.

In the SymDPO dataset, we employ a more challenging approach: firstly, all answers in the demonstrations are replaced with semantically meaningless symbols. This transformation can be expressed as:

$$\dot{D}_i = \{I_i, Q_i, S_i\}, \quad i \in \{1, 2, \ldots, N\} \tag{1}$$

where S_i represents a symbol unrelated to the actual answer, effectively stripping away semantic information to prevent the model from deducing the answer solely through simple textual patterns. This design compels the model to rely on a combination of visual and textual information within the ICDs to respond accurately within a multimodal environment.

Furthermore, a unique demonstration D_k is designated within the ICDs, in which the symbolic answer S_k aligns with the answer to the final question-answer pair $F = \{\hat{I}, \hat{Q}, \hat{A}\}$. For this setup, the chosen answer for F is set as S_k, while the rejected answer can be another unchosen answer, such as a different answer A_j from the same question type or another symbolic answer S_j, provided it satisfies the following conditions:

$$A_j \neq \hat{A} \quad \text{and} \quad S_j \neq S_k \tag{2}$$

For example, in the second data configuration of SymDPO in Figure 2, we replace "narrow" and "wide" with the symbols "rhondda" and "odwyer". In this scenario, even if the model can reason independently of the ICD, it must still interpret the overall semantics of these symbols within the ICD to answer correctly. This approach ensures that the model needs to understand the implicit meaning of the symbols deeply, rather than relying solely on isolated textual or visual information.

Finally, we add further complexity by introducing additional configurations, such as whether to erase the question in the ICD or whether the Rejected Answer should be semantically relevant. These adjustments lead to five different types of SymDPO data, maintaining a certain proportion of representation across all types in the final dataset to maximize diversity.

3.2. SymDPO

In the previous section, we introduced the data construction process of SymDPO. In this section, we will explain the principles behind SymDPO. Training LMMs with SymDPO can make LMMs pay more attention to visual information. Aligning preferences in Large Multimodal Models (LMMs) involves aligning the model's preferences with human preferences, typically employing techniques such as RLHF[37] (Reinforcement Learning from Human Feedback) and RLAIF[8] (Reinforcement Learning from AI Feedback). Considering a dataset $\mathcal{D}_S$, each entry includes an input $x = \{q, I, C\}$, a chosen response y_w and a rejected response y_l, while q represents the question, I represent images and C represents the context. $\mathcal{D}_S$ can be represented as $\mathcal{D}_S = \{x, y_w, y_l\}$.

Upon receiving the input x, a LMM produces an output y, to which a reward $r(x, y)$ is allocated. The reward model assesses both chosen (high $r(x, y)$) and rejected (low $r(x, y)$) samples. Meanwhile, to avoid overfitting on the dataset $\mathcal{D}_S$, preference alignment in LMMs incorporates a Kullback-Leibler (KL) divergence loss D_{KL}, which normalizes the disparity between the model's policy $\pi_\theta(y|x)$ and the reference model's policy $\pi_{ref}(y|x)$. The goal is to maximize this:

$$\max_\theta \left[\mathbb{E}_{x\sim\mathcal{D}_S,(y)\sim\pi_\theta(y|x)}[r(x,y)] - \beta \cdot D_{KL}(\pi_\theta(y|x)||\pi_{ref}(y|x))\right] \tag{3}$$

Here, θ denote the parameters of the LMM, π_θ denote the policy of the LMM, $\pi_\theta(y|x)$ denote the distribution of the LMM and the hyperparameter β controls the impact of the KL divergence within the optimization target. The reference model is the model's state prior to preference alignment.

To enhance the preference alignment target, the Direct Preference Optimization (DPO) method is utilized. The DPO method is efficient, stable, and does not require fitting a reward model. Our method, SymDPO, is based on the classical DPO algorithm. The SymDPO objective is formally defined as follows:

$$\begin{aligned}&\mathcal{L}_S(\pi_\theta; \pi_{ref}) = \\ &-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_S} log\sigma(\beta log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)})\end{aligned} \tag{4}$$

where σ is the logistic function.

4. Experiment

4.1. Experiment Setting

Implementation Details: We instantiate our model with Open-Flamingo [6] and IDEFICS [38], constructing the

Model	Shots	Method	COCO Caption (CIDEr)	Flickr-30K (CIDEr)	VQAv2 (Acc)	OK-VQA (Acc)	TextVQA (Acc)
OF-3B (I)	4	Base	82.7	59.1	45.7	30.6	28.1
		+ SymDPO	**87.4**$_{+4.7}$	**61.2**$_{+2.1}$	**46.2**$_{+0.5}$	**31.6**$_{+1.0}$	28.3$_{+0.2}$
		+ General DPO	83.5$_{+0.8}$	60.0$_{+0.9}$	46.0$_{+0.3}$	30.7$_{+0.1}$	28.2$_{+0.1}$
		+ Video DPO	82.5$_{-0.2}$	59.5$_{+0.4}$	45.5$_{-0.2}$	30.3$_{-0.3}$	28.4$_{+0.3}$
		+ MIA-DPO	84.7$_{+2.0}$	60.8$_{+1.7}$	46.1$_{+0.4}$	30.4$_{-0.2}$	**28.5**$_{+0.4}$
	8	Base	87.8	60.7	45.9	31.5	29.1
		+ SymDPO	**91.2**$_{+3.4}$	**65.3**$_{+4.6}$	**46.5**$_{+0.6}$	**32.7**$_{+1.2}$	**29.8**$_{+0.7}$
		+ General DPO	88.4$_{+0.6}$	61.5$_{+0.8}$	46.1$_{+0.2}$	31.3$_{-0.2}$	29.1$_{+0.0}$
		+ Video DPO	87.3$_{-0.5}$	60.6$_{-0.1}$	46.0$_{+0.1}$	31.4$_{-0.1}$	28.7$_{-0.4}$
		+ MIA-DPO	89.0$_{+1.2}$	62.5$_{+1.8}$	46.3$_{+0.4}$	31.2$_{-0.3}$	29.3$_{+0.2}$
	16	Base	91.9	63.0	45.8	30.7	29.1
		+ SymDPO	**93.4**$_{+1.5}$	**66.1**$_{+3.1}$	**46.5**$_{+0.7}$	**32.8**$_{+1.9}$	**29.6**$_{+0.5}$
		+ General DPO	92.0$_{+0.1}$	62.7$_{-0.3}$	46.0$_{+0.2}$	30.5$_{-0.2}$	29.0$_{-0.1}$
		+ Video DPO	91.8$_{-0.1}$	62.8$_{-0.2}$	45.9$_{+0.1}$	30.9$_{+0.2}$	29.2$_{+0.1}$
		+ MIA-DPO	92.5$_{+0.6}$	63.2$_{+0.2}$	46.1$_{+0.3}$	31.1$_{+0.4}$	29.4$_{+0.3}$
OF-9B	4	Base	89.0	65.8	54.8	40.1	28.2
		+ SymDPO	**93.8**$_{+4.8}$	**69.4**$_{+3.6}$	**56.8**$_{+2.0}$	**41.0**$_{+0.9}$	**28.8**$_{+0.6}$
		+ General DPO	89.2$_{+0.2}$	66.4$_{+0.6}$	55.2$_{+0.4}$	40.3$_{+0.2}$	28.5$_{+0.3}$
		+ Video DPO	88.7$_{-0.3}$	65.7$_{-0.1}$	54.7$_{-0.1}$	40.5$_{+0.4}$	28.7$_{+0.5}$
		+ MIA-DPO	88.6$_{-0.4}$	67.5$_{+1.7}$	55.2$_{+0.4}$	40.7$_{+0.6}$	28.9$_{+0.7}$
	8	Base	96.3	62.9	54.8	41.1	29.1
		+ SymDPO	**102.5**$_{+6.2}$	**67.3**$_{+4.4}$	**55.6**$_{+0.8}$	**42.3**$_{+1.2}$	**30.1**$_{+1.0}$
		+ General DPO	96.5$_{+0.2}$	62.9$_{+0.0}$	55.0$_{+0.2}$	41.5$_{+0.4}$	29.3$_{+0.2}$
		+ Video DPO	95.7$_{-0.6}$	62.8$_{-0.1}$	55.1$_{+0.3}$	40.2$_{-0.9}$	29.0$_{-0.1}$
		+ MIA-DPO	97.0$_{+0.7}$	63.5$_{+0.6}$	55.3$_{+0.5}$	40.2$_{-0.9}$	29.7$_{+0.6}$
	16	Base	98.8	62.8	54.3	42.7	27.3
		+ SymDPO	**104.3**$_{+5.5}$	**64.9**$_{+2.1}$	**55.7**$_{+1.4}$	**44.5**$_{+1.8}$	**28.1**$_{+0.8}$
		+ General DPO	98.9$_{+0.1}$	63.0$_{+0.2}$	54.5$_{+0.2}$	41.9$_{-0.8}$	27.5$_{+0.2}$
		+ Video DPO	98.2$_{-0.6}$	62.2$_{-0.6}$	54.6$_{+0.3}$	42.7$_{+0.0}$	26.7$_{-0.6}$
		+ MIA-DPO	98.5$_{-0.3}$	62.9$_{+0.1}$	54.8$_{+0.5}$	43.1$_{+0.4}$	26.9$_{-0.4}$
IDEFICS-9B	4	Base	93.0	59.7	55.4	45.4	27.6
		+ SymDPO	**96.5**$_{+3.5}$	**64.0**$_{+4.3}$	**56.1**$_{+0.7}$	**47.2**$_{+1.8}$	**28.6**$_{+1.0}$
		+ General DPO	93.2$_{+0.2}$	60.2$_{+0.5}$	55.6$_{+0.2}$	45.9$_{+0.5}$	27.8$_{+0.2}$
		+ Video DPO	93.5$_{+0.5}$	59.6$_{-0.1}$	55.7$_{+0.3}$	45.8$_{+0.4}$	28.0$_{+0.4}$
		+ MIA-DPO	93.7$_{+0.7}$	61.5$_{+1.8}$	55.9$_{+0.5}$	46.3$_{+0.9}$	28.2$_{+0.6}$
	8	Base	97.0	61.9	56.4	47.7	27.5
		+ SymDPO	**103.8**$_{+6.8}$	**66.1**$_{+4.2}$	**57.2**$_{+0.8}$	**49.5**$_{+1.8}$	**28.5**$_{+1.0}$
		+ General DPO	97.2$_{+0.2}$	62.0$_{+0.1}$	56.6$_{+0.2}$	48.1$_{+0.4}$	27.7$_{+0.2}$
		+ Video DPO	97.5$_{+0.5}$	62.2$_{+0.3}$	56.2$_{-0.2}$	47.3$_{-0.4}$	27.9$_{+0.4}$
		+ MIA-DPO	97.7$_{+0.7}$	62.5$_{+0.6}$	56.9$_{+0.5}$	48.3$_{+0.6}$	28.1$_{+0.6}$
	16	Base	99.7	64.5	57.0	48.4	27.9
		+ SymDPO	**107.9**$_{+8.2}$	**69.3**$_{+4.8}$	**58.2**$_{+1.2}$	**50.6**$_{+2.2}$	**29.3**$_{+1.4}$
		+ General DPO	99.6$_{-0.1}$	64.7$_{+0.2}$	57.2$_{+0.2}$	43.8$_{-4.6}$	28.8$_{+0.9}$
		+ Video DPO	99.4$_{-0.3}$	63.9$_{-0.6}$	57.3$_{+0.3}$	48.3$_{-0.1}$	28.0$_{+0.1}$
		+ MIA-DPO	99.8$_{+0.1}$	64.0$_{-0.5}$	57.5$_{+0.5}$	48.2$_{-0.2}$	28.2$_{+0.3}$

Table 1. Comparison of Different DPO Methods: Performance of models with various DPO techniques across benchmarks and shot counts.

SymDPO dataset from the VQAv2 [5], GQA [20], and ImageNet [13] training sets, amassing a total of 872,000 data items. From this collection, a subset of 10,000 samples is randomly selected for training. To enhance data quality, we apply GPT-4v to the selected samples. In the post-training phase, we employ linear annealing to adjust the learning rate, initializing it at 5e-6. This task is executed on 8 NVIDIA A100 GPUs, with the complete post-training

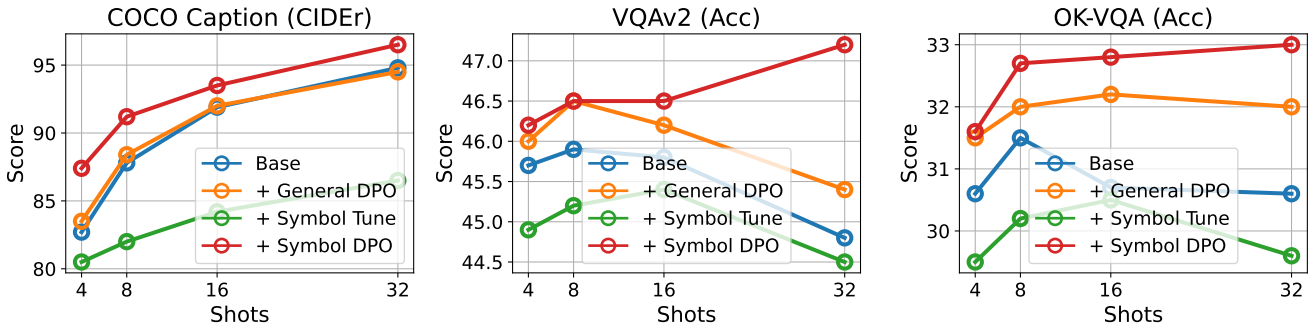

Figure 3. Comparison of Symbol Tuning, General DPO, and SymDPO Methods: We optimized OF 3b using three different methods: Symbol Tuning, General DPO, and SymDPO, resulting in three distinct variants. The performance of these variants was visualized using line charts, showcasing the results across four-shot (4, 8, 16, 32) settings on the COCO, VQAv2, and OK-VQA benchmarks.

process taking approximately 1 hour.

Benchmark: Our model is evaluated in alignment with Flamingo on image captioning benchmarks, speci fically COCO Caption [12] and Flickr 30K [50], as well as on three question-answering (QA) benchmarks: VQA v2 [5], OK-VQA [33], and TextVQA [41]. For the image captioning tasks, we report CIDEr scores [44] as evaluation metrics, while for QA tasks, we use accuracy (Acc) as the metric.

Baseline: We compare SymDPO against two different DPO optimization approaches:

- Video DPO [53] - Proposed by LLaVA-Hound-DPO, this approach utilizes a video-specific DPO dataset to improve the model's understanding of video data.
- MIA-DPO [31] - Designed for multi-image scenarios, this method aims to mitigate hallucinations in LMMs by optimizing in multi-image settings.

4.2. Main Results

As illustrated in Table 1, we evaluated the performance of SymDPO, Video DPO, and MIA-DPO on Open-Flamingo (OF) and IDEFICS 9B across five different benchmarks. The results reveal that SymDPO consistently enhances performance across all benchmarks for both OF and IDEFICS, demonstrating the efficacy of SymDPO. In contrast, Video DPO showed no notable improvement, while MIA-DPO yielded only marginal gains. We attribute these outcomes to the specific design focuses of Video DPO and MIA-DPO: Video DPO is primarily oriented toward semantic alignment and optimization for video data, whereas MIA-DPO targets alignment for generic multi-image instructions. Neither approach explicitly addresses the instruction alignment in in-context scenarios, a key focus of SymDPO. We interpret this as a result of the expanded contextual knowledge composed of both visual and textual elements, allowing an LMM fine-tuned with SymDPO to better integrate and leverage this combined knowledge, thus achieving greater performance gains.

4.3. Ablation Study

4.3.1. Effectiveness of SymDPO

To further assess the effectiveness of SymDPO, we conducted several ablation experiments.

General DPO vs. SymDPO: As outlined in Method Subsection 3.1, the General DPO approach employs a standard DPO dataset for optimization, without replacing answers with symbols as SymDPO does. As shown in Table 1 and Figure 3, we observed that models optimized with General DPO (i.e., Open-Flamingo (OF) and IDEFICS) exhibit significantly lower performance improvements compared to those optimized with SymDPO. This result substantiates the advantage of the symbolic answer replacement strategy within SymDPO, affirming its effectiveness.

Visual Context Overlook Investigation: To determine whether SymDPO's enhancements arise from addressing the visual context overlook issue in large multimodal models, we conducted additional tests. Specifically, we modified the demonstration data by either replacing images with blank placeholders ("w/ blank") or omitting images altogether ("w/o image"). We then evaluated the performance of OF and OF+SymDPO ("OF-s") under these modified conditions. The results, displayed in Figure 4, reveal that the performance of OF-SymDPO significantly declines when images are removed, suggesting that the model's advantage derives from its comprehensive understanding of both visual and textual information in in-context demonstrations, rather than relying solely on textual data. This further emphasizes SymDPO's capability in leveraging the integrated visual-textual knowledge, enabling a more robust and contextually aware model.

Symbol DPO VS. Symbol Finetuning: To validate the advantages of DPO-based optimization, we conducted additional experiments following the preference data collection methodology in Method Subsection 3.1. After collecting the preference dataset, we constructed a multimodal symbolic fine-tuning dataset, inspired by the Symbol Tuning approach [47]. In this setup, we used the chosen answer as the target label for an autoregressive generation task during

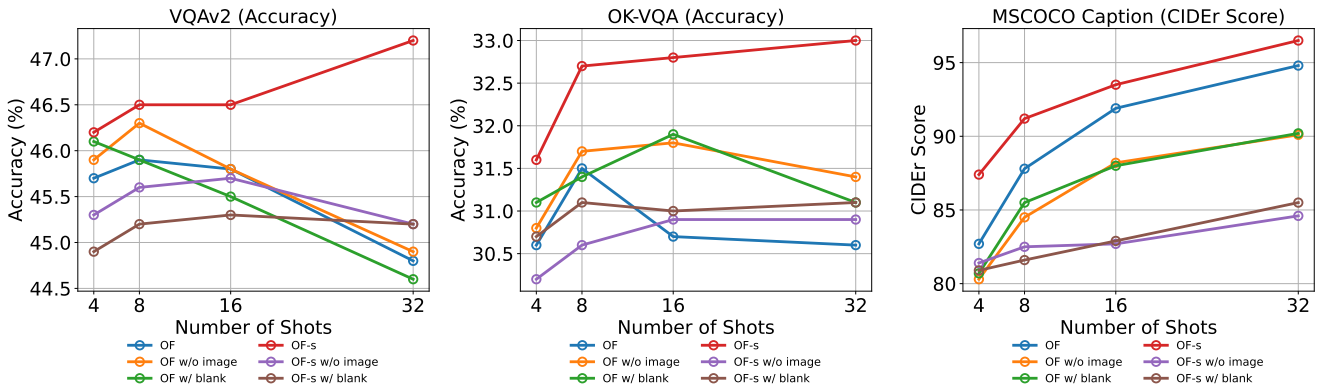

Figure 4. Impact of Visual Context Removal on OF and OF+SymDPO Performance.

Model	Shots	Method	COCO Caption (CIDEr)	Flickr-30K (CIDEr)	VQAv2 (Acc)	OK-VQA (Acc)	TextVQA (Acc)
OF-3B (I)	4	Base	82.7	59.1	45.7	30.6	28.1
		+ RICES	90.5$_{+7.8}$	53.9$_{-5.2}$	45.3$_{-0.4}$	31.4$_{+0.8}$	28.9$_{+0.8}$
		+ SymDPO	87.4$_{+4.7}$	61.2$_{+2.1}$	45.8$_{+0.1}$	31.6$_{+1.0}$	28.3$_{+0.2}$
		+ SymDPO & RICES	**93.5**$_{+10.8}$	**62.0**$_{+2.9}$	**46.6**$_{+0.9}$	**33.4**$_{+2.8}$	**29.1**$_{+1.0}$
	8	Base	87.8	60.7	45.9	31.5	29.1
		+ RICES	96.8$_{+9.0}$	58.6$_{-2.1}$	46.1$_{+0.2}$	32.8$_{+1.3}$	28.8$_{-0.3}$
		+ SymDPO	91.2$_{+3.4}$	65.3$_{+4.6}$	46.5$_{+0.6}$	32.7$_{+1.2}$	29.8$_{+0.7}$
		+ SymDPO & RICES	**98.4**$_{+10.6}$	**68.2**$_{+7.5}$	**47.2**$_{+1.3}$	**34.3**$_{+2.8}$	**31.7**$_{+2.6}$
	16	Base	91.9	63.0	45.8	30.7	29.1
		+ RICES	101.1$_{+9.2}$	61.5$_{-1.5}$	46.6$_{+0.8}$	33.9$_{+3.2}$	28.8$_{-0.3}$
		+ SymDPO	93.4$_{+1.5}$	64.6$_{+1.6}$	46.5$_{+0.7}$	32.8$_{+1.9}$	29.6$_{+0.5}$
		+ SymDPO & RICES	**106.8**$_{+14.9}$	**66.5**$_{+3.5}$	**47.2**$_{+1.4}$	**35.1**$_{+4.4}$	**29.8**$_{+0.7}$

Table 2. Performance comparison of the OF-3B (I) model using RICES and SymDPO across various datasets and shot counts.

model fine-tuning, producing the variant OF3B + SymTune. The experimental results, as shown in Table 3, indicate that SymTune does not achieve satisfactory outcomes; notably, its performance on captioning tasks even declines. In contrast, the performance gains of OF3B + SymDPO are substantial across all benchmarks. We attribute this difference to the following key factors: The SymTune approach relies on symbolic fine-tuning where the model learns to predict the chosen answer directly in an autoregressive manner. However, this approach may not fully exploit preference data's structured feedback, resulting in limited guidance for multimodal alignment.

4.3.2. Effect of Different Demonstration Selection Strategies

In the ICL scenario, the choice of demonstration examples (demos) can significantly influence the reasoning performance of large multimodal models. However, prior research has also noted that, due to the issue of visual context overlooks, varying the demo selection does not markedly impact LMM performance. To validate the effectiveness of SymDPO, we employed the RICES (Retrieval In-Context Example Selection) method, as used in Flamingo, to select demos. We then re-evaluated the performance of OpenFlamingo 3B (OF) and OF + SymDPO across different benchmarks. As shown in Table 2, introducing RICES leads to a more pronounced improvement in the SymDPO model. This finding further highlights that incorporating SymDPO enables LMMs to leverage the integrated semantics and knowledge present in the demo's visual-textual content more effectively.

4.3.3. SymDPO and General DPO Integration: Impact on Task Performance

As depicted in Figure 5, we investigated the effects of integrating SymDPO and General DPO at varying proportions on task-specific performance throughout the alignment phase of DPO optimization. Symbolic data was incrementally introduced into the OF optimization process at ratios of 0%, 30%, 50%, 70%, and 100%, and model performance was subsequently evaluated across benchmark datasets. The experimental results presented in Figure 5 indicate that model performance improves with increasing proportions of symbolic data, particularly within question-answering (QA) tasks. However, an exclusive reliance on SymDPO data does not yield optimal performance. Our

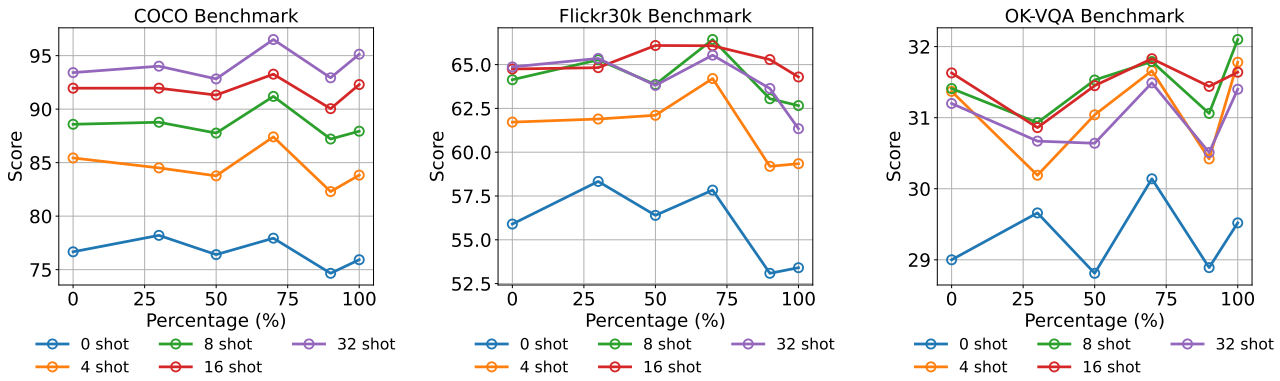

Figure 5. Comparison of the Impact of General DPO and SymDPO on LMMs with Varying Data Proportions in the Preference Dataset

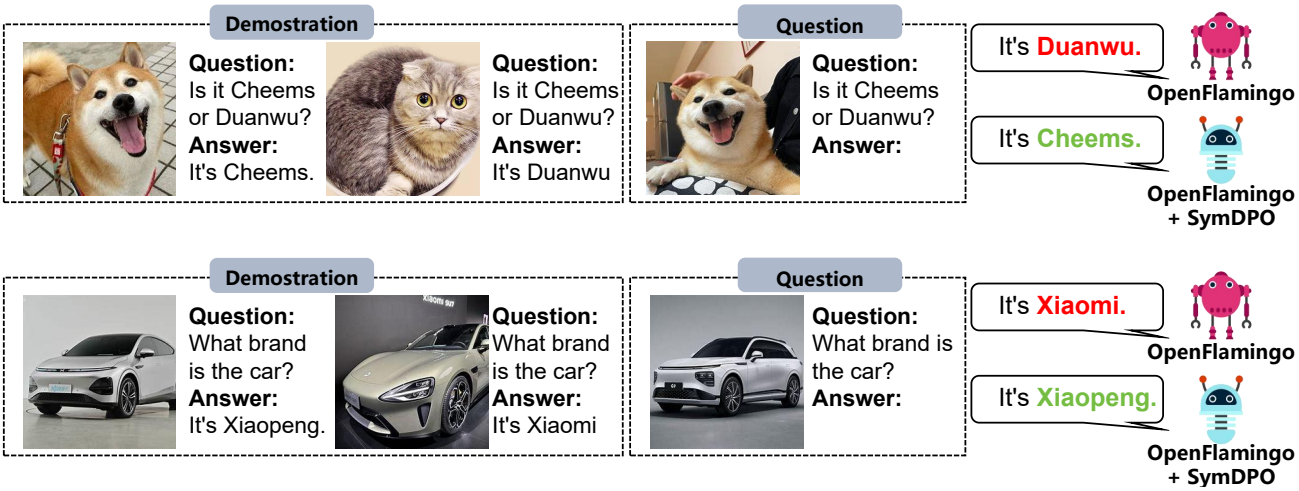

Figure 6. Example Visualization of OpenFlamingo 3B and OpenFlamingo 3B + SymDPO in ICL 2-Shot Setting.

findings show that a 70% symbolic data ratio achieves peak effectiveness; whereas a 100% symbolic data ratio is more effective for the OK-VQA task, suggesting task-specific dependencies on the symbolic data ratio.

4.4. Case Study

For further quantitative analysis of our method's effectiveness, we visualized several diverse In-Context Learning (ICL) scenarios. As illustrated in Figure 6, OF+SymDPO consistently yields accurate answers by interpreting the semantic context within demonstrations, whereas OF alone often misinterprets the task, relying predominantly on proximate textual information rather than fully understanding the demonstration content. In the first case, a demonstration provides the names of two pets: a dog named "Cheems" and a cat named "Duanwu." When the model is shown an image of the dog and asked to identify its name, OF+SymDPO accurately answers "Cheems," whereas OF responds incorrectly with "Duanwu," influenced by nearby textual cues without integrating the visual context. This pattern recurs across other cases, indicating that SymDPO effectively addresses the "visual context overlook" issue in Large Multimodal Models (LMMs). By doing so, SymDPO enables these models to comprehend and utilize both visual and textual information from demonstrations in a more holistic manner.

5. Conclusion

This work presented SymDPO, a symbolic preference optimization method designed to tackle the visual context overlooked in LMMs. By enforcing reliance on both visual and textual cues in In-Context Demonstrations, SymDPO effectively reduces LMMs' tendency toward textual pattern matching. Experiments confirm that SymDPO improves multimodal comprehension by compelling models to integrate visual context meaningfully, leading to consistent performance gains across benchmarks. In sum, SymDPO provides a robust approach to enhancing multimodal learning, marking a step toward more contextually aware LMMs.

References

[1] Gpt-4v(ision) system card. 2023. 1

[2] Jacob D. Abernethy, Alekh Agarwal, Teodor Vanislavov Marinov, and Manfred K. Warmuth. A mechanism for sample-efficient in-context learning for sparse retrieval tasks. *ArXiv*, abs/2305.17040, 2023. 2

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

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 5, 6

[6] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390, 2023. 1, 4

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. 4

[9] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1539–1550, 2024. 2

[10] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1, 2

[11] Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. *ArXiv*, abs/2311.18021, 2023. 2

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[14] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023. 3

[15] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 1, 2

[16] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2

[17] Sivan Doveh, Shaked Perek, Muhammad Jehanzeb Mirza, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision & language models. *ArXiv*, abs/2403.12736, 2024. 2

[18] Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. Structured prompting: Scaling in-context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*, 2022. 2

[19] Or Honovich, Uri Shaham, Samuel R Bowman, and Omer Levy. Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*, 2022. 2

[20] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[21] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 2

[22] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *ArXiv*, abs/2312.14925, 2023. 3

[23] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[24] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 1

[25] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, 2023. 3

[26] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 1, 2

[27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model

with in-context instruction tuning. *ArXiv*, abs/2305.03726, 2023. 2

[28] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *ArXiv*, abs/2205.05638, 2022. 2

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2

[30] Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables and images. *ArXiv*, abs/2309.04790, 2023. 2

[31] Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*, 2024. 6

[32] Ziyu Liu, Yuhang Zang, Xiao wen Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. 2024. 2

[33] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 6

[34] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*, 2021. 2

[35] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*, 2023. 2

[36] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, Tom Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. In-context learning and induction heads. *ArXiv*, abs/2209.11895, 2022. 2

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 1, 3, 4

[38] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *ArXiv*, abs/2403.08730, 2024. 2, 4

[39] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023. 2, 3

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. 3

[41] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 6

[42] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14398–14409, 2023. 1

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[44] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[45] Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. *ArXiv*, abs/2406.11839, 2024. 2

[46] Yifan Wang, Qingyan Guo, Xinzhe Ni, Chufan Shi, Lemao Liu, Haiyun Jiang, and Yujiu Yang. Hint-enhanced in-context learning wakes large language models up for knowledge-intensive tasks. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10276–10280, 2023. 2

[47] Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*, 2023. 6

[48] Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. *ArXiv*, abs/2303.07895, 2023. 2

[49] Xu Yang, Yingzhe Peng, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, and Hanwang Zhang. Lever lm: Configuring in-context sequence to lever large vision language models. *arXiv e-prints*, pages arXiv–2312, 2023. 2

[50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[51] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023. 3

[52] Chaoyi Zhang, Kevin Qinghong Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and

Lijuan Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13647–13657, 2023. 2

[53] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024. 6

[54] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *ArXiv*, abs/2309.07915, 2023. 2