ISR-DPO: Aligning Large Multimodal Models for Videos by Iterative Self-Retrospective DPO

Daechul Ahn[1*], **Yura Choi**[1,2*], **San Kim**[1], **Youngjae Yu**[2], **Dongyeop Kang**[3], **Jonghyun Choi**[1†]

[1]Seoul National University [2]Yonsei University [3]University of Minnesota

{daechulahn,00sankim,jonghyunchoi}@snu.ac.kr {yoorachoi,yjy}@yonsei.ac.kr dongyeop@umn.edu

Abstract

Iterative self-improvement, a concept extending beyond personal growth, has found powerful applications in machine learning, particularly in transforming weak models into strong ones. While recent advances in natural language processing have shown its efficacy through iterative preference optimization, applying this approach to Video Large Multimodal Models (VLMMs) remains challenging due to modality misalignment. VLMMs struggle with this misalignment during iterative preference modeling, as the self-judge model often prioritizes linguistic knowledge over visual information. Additionally, iterative preference optimization can lead to visually hallucinated verbose responses due to length bias within the self-rewarding cycle. To address these issues, we propose Iterative Self-Retrospective Direct Preference Optimization (ISR-DPO), a method that uses self-retrospection to enhance preference modeling. This approach enhances the self-judge's focus on informative video regions, resulting in more visually grounded preferences. In extensive empirical evaluations across diverse video question answering benchmarks, the ISR-DPO significantly outperforms the state of the art. We are committed to open-sourcing our code, models, and datasets to encourage further investigation. https://github.com/snumprlab/ISR-DPO

1 Introduction

Progress is not achieved by luck or accident, but by working on yourself daily.

— Epictetus

The human capacity for growth through consistent effort and repetition is a fundamental principle of personal development (Dweck 2006). This concept of iterative self-improvement extends beyond personal growth, finding powerful applications in machine learning to transform weak models into strong ones, without relying on additional human-annotated training data (Schapire 1990; Yuan et al. 2024; Burns et al. 2023). Notably, recent advances in natural language processing (NLP) have demonstrated the efficacy of *iterative* preference optimization in aligning Large Language Models (LLMs) with human intentions (Yuan et al. 2024; Pang et al. 2024; Chen et al. 2024). This approach involves constructing increasingly informative preferences through iterative preference modeling, *i.e.*, LLM-as-a-judge, leading to progressively better-aligned models.

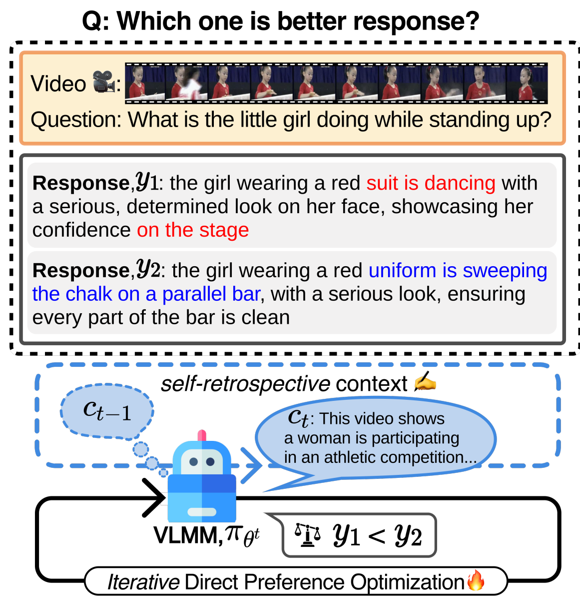

Figure 1: **Illustration of the proposed ISR-DPO.** During iterative direct preference optimization (DPO) in VLMM, we select preferences from responses based on not only video content but also visual context c_t, *i.e.*, detailed video description, to ensure preferences are grounded in video information. Specifically, we enhance the context in the self-retrospective manner by leveraging context c_{t-1} generated in previous iteration, a process we call *self-retrospective* preference modeling. Red indicates irrelevant responses, while blue indicates accurate, visually-grounded responses.

However, this iterative self-improvement principle for LLMs poses specific challenges when applied to large multimodal models, particularly Video Large Multimodal Models (VLMMs). VLMMs suffer from modality misalignment during iterative preference modeling, where the self-judge

---
*These authors contributed equally.

†JC is with ECE, ASRI and IPAI in SNU and a corresponding author.

Question: Why does the video suggest using lip gloss instead of lipstick?

Ground-Truth : A text overlay in the video says, “Lipgloss is easier to touch up than lipstick!”, emphasizing the convenience of using lip gloss.

#1st iteration

The video suggests using lip gloss instead of lipstick because it is implied to be easier to apply and less messy.

#9th iteration

The video suggests using lip gloss instead of lipstick because it is implied to be easier to apply and less messy. The text overlays in the video explicitly state that "Lip gloss is easier to touch up than lipstick!" This implies that lip gloss offers a more convenient and accessible beauty option, especially for on-the-go touch-ups. Additionally, the video advises to "Go with a lip balm on a tube so it doesn't get everywhere!" which further emphasizes the practicality and ease of use of lip gloss compared to lipstick. The video seems to be promoting the idea that lip gloss is a practical and user-friendly beauty product.

Figure 2: **Example of verbosity hallucination within iterative preference modeling cycle for VLMM.** At the 1st iteration, the response is concise and visually grounded (in blue). By the $9th$ iteration, the response elaborates further, referencing explicit text overlays in the video. However, it starts to include irrelevant details and assumptions as well, leading to *verbosity hallucination* highlighted in red.

model tends to rely more on their pre-existing linguistic knowledge rather than the given visual information (Ahn et al. 2024; Zhou et al. 2024). This leads to preference data that are linguistically plausible but less grounded in visual content. Moreover, iterative training exacerbates the visually *ungrounded* verbose response in VLMMs due to the length bias within the iterative preference modeling cycle, which favors linguistically *longer* response during preference selection (Prasann Singhal and Durrett 2023; Park et al. 2024). As illustrated in Fig. 2, while somewhat longer responses might enhance the quality of the predicted response, excessively long responses can introduce content irrelevant to the actual video or question, *i.e.*, *verbosity hallucination*, without necessarily improving quality.

To address these challenges, we argue that the self-judge model, *i.e.*, VLMM, should select preferences based on visual content, rather than being merely linguistically plausible at each iteration. We achieve this visually grounded self-judgment by drawing inspiration from cognitive science on human perception (Bransford and Johnson 1972; Kintsch 1988; Anderson 1984), emphasizing the importance of contextual information in interpreting visual data. Specifically, we provide the self-judge with additional video descriptions generated through a self-retrospective manner as an additional visual context. This additional information acts as a focusing mechanism, akin to attention in human cognition (Bransford and Johnson 1972), enabling the VLMM to ground its responses more effectively in the video, reducing the likelihood of generating irrelevant or hallucinated one.

To this end, we propose a simple yet effective iterative self-improvement approach for VLMM: **I**terative **S**elf-**R**etrospective **D**irect **P**reference **O**ptimization (ISR-DPO) as shown in Fig. 1. This approach helps the self-judge focus on more informative regions in the video when comparing responses, producing more visually grounded preferences at each iteration. Our empirical studies demonstrate that our ISR-DPO exhibits superior performance compared to state-of-the-art VLMMs on various video question answering benchmarks.

We summarize our contributions as follows:

- We propose a novel modality alignment method for video large multimodal models (VLMMs), utilizing iterative direct preference optimization (DPO) to align video-text modalities effectively.
- We enhance AI's feedback by proposing self-retrospective preference modeling, which improves clarity and comprehension in video through the use of iteratively refined visual context for preference selection.
- We demonstrate the effectiveness of our proposed ISR-DPO on various video question answering benchmarks by a noticeable margin.

2 Related Work

Aligning large multimodal models for videos. VLMMs have achieved notable success in various video comprehension tasks, such as video temporal understanding (Liu et al. 2023), question answering (Lin et al. 2023), and instruction-following (Maaz et al. 2024). These models integrate publicly available LLMs (Touvron et al. 2023a,b) with visual encoders (Radford et al. 2021) and additional learnable parameters (Hu et al. 2022), undergoing Supervised Fine-Tuning (SFT) (Maaz et al. 2024; Lin et al. 2023; Li, Wang, and Jia 2023) and, more recently, preference optimization (Rafailov et al. 2023; Zhang et al. 2024a; Ahn et al. 2024). Our work builds upon these efforts by exploring the application of iterative preference optimization to VLMMs and addressing the unique challenges related to length bias and visual grounding during preference modeling process.

Iterative preference optimization. Training LLMs with preference optimization has proven to be an effective approach to align language models with human intention, improving model performance and reliability. Build upon this preference optimization, recent efforts have focused on iterative preference optimization techniques, which typically involve iteratively generating feedback data with AI models themselves, *i.e.*, *self-rewarding*. Many recent work in the NLP domain concurrently propose this, where the aligned model iteratively generates responses and judges its own outputs to build feedback data and learn from this data with DPO (Yuan et al. 2024; Pang et al. 2024; Chen et al. 2024). While these iterative optimization techniques have shown their effectiveness in LLMs, their application in the multimodal domain, particularly for video understanding tasks, remains largely unexplored. Our work proposes an effective iterative preference optimization method for VLMMs.

Verbosity bias in preference optimization. Preference fine-tuning methods such as RLHF, RLAIF, and DPO are

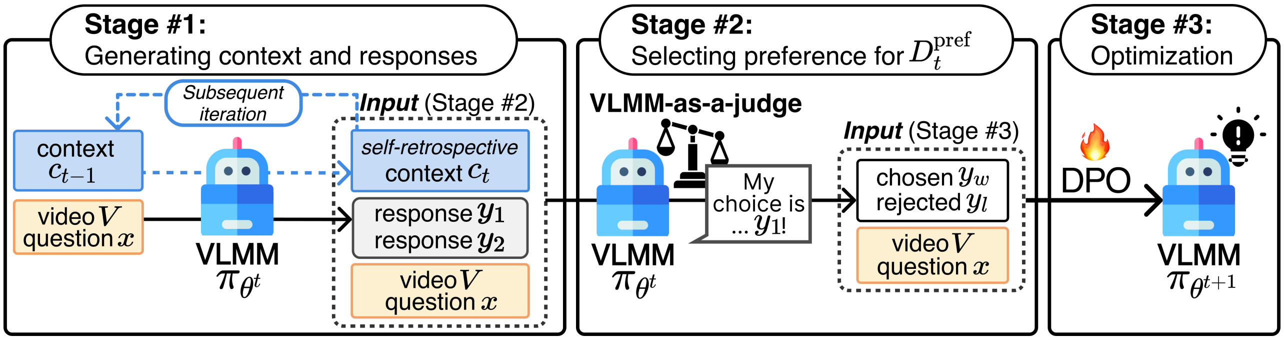

Figure 3: **Overview of self-retrospective Direct Preference Optimization (DPO).** Each iteration of ISR-DPO involves three stages: 1) After training iteration t, the latest updated VLMM (π_{θ^t}) generates two different responses y_1 and y_2 for the given video V and instruction x. In addition, a visual description, *i.e.*, visual context, is generated through self-retrospection, providing the necessary input for the next stage, as indicated by the black dotted line. 2) Using the information generated in the previous stage, the model (π_{θ^t}) compares its responses(y_1 and y_2) and classifies the preferred response y_w and the rejected response y_l. 3) Then, the VLMM (π_{θ^t}) is optimized using DPO to update the parameters to $\pi_{\theta^{t+1}}$.

known to produce responses that are longer than those generated prior to preference optimization, known as length bias. This phenomenon stems from a verbosity bias in preference data, where both human and AI judges tend to favor longer responses (Prasann Singhal and Durrett 2023; Park et al. 2024; Saito et al. 2023). Despite minimal differences in length between preferred and rejected responses, the increase in verbosity is statistically significant (Park et al. 2024). In VLMMs, this length bias can be particularly problematic. It may result in verbose responses that are linguistically comprehensible but not well-grounded in the visual content. Addressing length bias in the multimodal setting of VLMMs remains an open challenge.

3 Iterative Self-Retrospective DPO

To effectively align the multimodalities between video and text, we propose to use an iterative self-improvement approach for VLMM. Figure 3 illustrates the overall training pipeline of our proposed ISR-DPO for one cycle, which executes three stages: 1) generating self-retrospective context and responses, 2) selecting preferences, and 3) optimization.

During iterative execution, we enhance our model's ability to select preferences by conditioning not only the video content, but also on the visual context generated through self-retrospection. This additional visual context generates preferences grounded in the video, improving the alignment between visual and textual modalities.

3.1 Iterative DPO in VLMM

We denote the current VLMM at the t-th iteration as π_{θ^t}. This model generates responses and selects preferences by itself, thereby constructing the preference data, D_t^{pref}. With D_t^{pref}, we train the subsequent VLMM, denoted as $\pi_{\theta^{t+1}}$, at the $t+1$-th iteration.

Initial model. Given a seed preference data annotated in Zhang et al. (2024a), we conduct preference fine-tuning using DPO, starting from the SFT model provided from previous work (Zhang et al. 2024a). This preference fine-tuned model is referred to as the initial model π_{θ^1}.

Preference modeling. Given the current VLMM π_{θ^t}, we generate two different responses for the input video V and question x using a high temperature hyper-parameter (*e.g.*, 0.7). This high temperature flattens the token sampling probability distribution, producing varied responses from the same input in the current VLMM π_{θ^t}:

$$y_1 \sim \pi_{\theta^t}(V, x),\ y_2 \sim \pi_{\theta^t}(V, x).$$

We then select a better response between two responses by leveraging the current VLMM to evaluate its own responses, *i.e.*, VLMM-as-a-judge. In particular, we provide the VLMM with the visual context c_t for enhanced visual clarity (more detailed in Sec. 3.2). We can present this preference selection procedure as follows:

$$(y_w, y_l) \sim \pi_{\theta^t}(V, x, c_t, y_1, y_2),$$

where y_1 and y_2 are two sampled responses, y_w is the chosen response, and y_l is the rejected response.

After constructing the preference data at t-th iteration as $D_t^{\text{pref}} = \{V, x, y_w, y_l\}$, we use this dataset to perform preference optimization on the current VLMM π_{θ^t} using DPO. The DPO objective for the current VLMM π_{θ^t} is represented as follows:

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta^t}; \pi_{\text{ref},t}) = - \mathbb{E}_{(V,x,y_w,y_l)\sim\mathcal{D}_{t-1}^{\text{pref}}} \Bigg[\log \sigma \Bigg(\beta \log \frac{\pi_{\theta^t}(y_w \mid V, x)}{\pi_{\text{ref},t}(y_w \mid V, x)} - \beta \log \frac{\pi_{\theta^t}(y_l \mid V, x)}{\pi_{\text{ref},t}(y_l \mid V, x)}\Bigg)\Bigg],$$

where $\pi_{ref,t}$ is the current base reference model, β is a hyper-parameter controlling the deviation from the current base reference model and σ is the sigmoid function.

Iterative training. Our overall iterative training procedure follows previous work (Yuan et al. 2024), where a series of models $\pi_{\theta^1}, \ldots, \pi_{\theta^T}$ is trained sequentially. Each successive model at iteration of $t+1$ uses preference data D_t^{pref} generated by the VLMM at iteration t, defined as follows:

$$\pi_{\theta^{t+1}} : \text{Training with } D_t^{\text{pref}} \text{ initialized from } \pi_{\theta^t},$$

where the t-th model π_{θ^t} generates preference data D_t^{pref} through self-judgment.

3.2 Self-Retrospective Preference Modeling

A key aspect of iterative DPO in VLMM involves using a VLMM as a judge to iteratively select preferences that accurately answer posed questions (Ahn et al. 2024). Specifically, we provide the VLMM with detailed visual descriptions as visual context, generated by the VLMM itself in addition to the video content for improved visual clarity. Moreover, inspired by humans learning process, we enhance the visual context in a *self-retrospective* manner. Just as retrospection allows humans to make better decisions by reflecting on the past (Simon 1962; Madaan et al. 2023), we leverage previously generated visual context to generate better context, enhancing the accuracy and relevance of the preference selection process, defined as follows:

$$c_t \sim \pi_{\theta^t}(V, c_{t-1}),$$

where c_{t-1} is the previous visual context at time $t-1$.

Using the generated context c_t, question x, video V, and responses $\{y_1, y_2\}$, we classify the chosen y_w and rejected data y_l from responses using the current aligned VLMM π_{θ^t}, a process we call *self-retrospective* preference modeling, thereby constructing preference data D_t^{pref} at time t.

4 Experiments

4.1 Experimental Setup

Dataset details. Our training dataset utilizes a fixed set of 17k video-instruction ($\{V, x\}$) pairs from (Zhang et al. 2024a), in contrast to previous works (Yuan et al. 2024; Chen et al. 2024) that incremented their dataset across iterations. For all iterations beyond the initial VLMM π_{θ^1}, we generate preference dataset D_t^{pref} at each iteration by generating new responses and preferences. Following (Maaz et al. 2024; Zhang et al. 2024a), we evaluate our method on two types of video question answering datasets: one that requires concise responses, and the other that demands comprehensive answers, across 7 video collections.

Training details. We perform full-parameter fine-tuning using DPO with 9 total iterations, tripling the previous iterative preference optimization approach for LLMs alignment (Yuan et al. 2024). All generative processes use specific prompts. Training is conducted on 8×NVIDIA A100 GPUs (80G). We employ a 7B-sized model for fair comparison with others.

4.2 Quantitative Analysis

In-domain video question answering. As shown in Tab. 1, ISR-DPO demonstrates consistent performance gains at each iteration up to the $9th$ iteration. Moreover, final

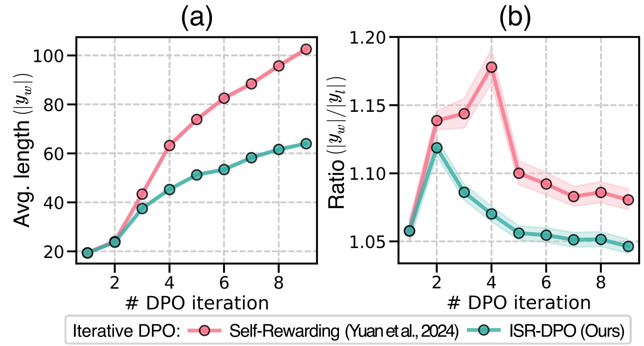

Figure 4: **Length analysis of preference dataset during iterative DPO.** (a) Average (Avg.) word length of chosen response $|y_w|$ in preference dataset D_t^{pref} across DPO iterations. Self-rewarding results in longer responses compared to the ISR-DPO. (b) Ratio of the word lengths of chosen responses ($|y_w|$) to rejected responses ($|y_l|$). ISR-DPO consistently maintains a lowered ratio compared to the self-rewarding, indicating reduced response length after optimized. '# DPO iteration' means the number of DPO iterations.

iterated model (π_{θ^9}), outperforms all previous work across all video benchmarks in both accuracy and score by a noticeable margin. We attribute this performance improvement to the better alignment of video modality provided by the proposed iterative retrospective judgment for VLMMs.

Out-domain video question answering. For evaluating out-domain video question answering, we use two types of datasets. Tables 2 and 3 show the comparative results for datasets that require complex answers and concise keyword answers, respectively. The final iterated model of ISR-DPO (π_{θ^9}) outperforms the previous work by a large margin in both cases, demonstrating its effectiveness in generating both detailed and precise responses. This model also shows consistent performance improvements at each iteration, as shown in Tables 2 and 3.

4.3 Detailed Analysis

To evaluate the effectiveness of ISR-DPO, we address the following research questions, specifically exploring the effect and design of visual context:

- **RQ1**: What are the effects and benefits of visual context during iterative DPO?
- **RQ2**: How should the visual context be designed?

In particular, we compare ISR-DPO with self-rewarding (Yuan et al. 2024), which serves as our baseline for adopting iterative DPO in VLMMs without self-retrospective context.

Effect of visual context during iterative process Figure 4 demonstrates the effect of including visual context during preference selection. As shown in Fig. 4-(a), ISR-DPO generates shorter chosen responses compared to self-rewarding as training iterations progress. Similarly, Fig. 4-(b) shows a lower ratio of chosen to rejected response

Methods	ActivityNet-QA		VIDAL-QA		WebVid-QA	
	Acc.	Score	Acc.	Score	Acc.	Score
Video-ChatGPT (Maaz et al. 2024)	34.17	2.19	29.35	2.10	38.88	2.27
LLaMA-VID (Li, Wang, and Jia 2023)	36.54	2.27	30.58	2.15	36.99	2.24
Chat-UniVi (Jin et al. 2023)	39.35	2.32	31.40	2.16	40.05	2.31
Video-LLaVA (Lin et al. 2023)	41.35	2.38	34.30	2.24	42.47	2.39
VLM-RLAIF† (Ahn et al. 2024)	53.27	2.56	44.82	2.40	53.69	2.62
PLLaVA† (Xu et al. 2024)	48.44	2.50	42.45	2.39	53.55	2.59
LLaVA-NeXT-DPO† (Zhang et al. 2024b)	68.05	2.88	61.52	2.72	73.35	3.00
LLaVA-Hound-DPO (Zhang et al. 2024a)	<u>76.62</u>	<u>3.18</u>	<u>70.06</u>	<u>3.04</u>	<u>79.82</u>	<u>3.29</u>
ISR-DPO (π_{θ^1})	75.58	3.14	70.07	3.02	80.74	3.28
ISR-DPO (π_{θ^5})	81.62	3.25	77.33	3.10	86.92	3.39
ISR-DPO (π_{θ^9})	**82.99**	**3.26**	**79.00**	**3.13**	**88.11**	**3.40**

Table 1: **Quantitative comparison between different VLMMs on *in-domain* video question answering with detailed captions as supporting evidence proposed in Zhang et al. (2024a).** Our final iterated model of ISR-DPO (π_{θ^9}) consistently outperforms all other models in both accuracy and score across these benchmarks, demonstrating superior performance in in-domain video question answering tasks. The best results are **bold** and the second-best results are <u>underlined</u>. †: reproduced by the authors' implementation. All results except † are directly sourced from Zhang et al. (2024a).

Methods	MSVD-QA		MSRVTT-QA		TGIF-QA		SSV2-QA	
	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
Video-ChatGPT (Maaz et al. 2024)	34.06	2.20	25.65	1.98	31.35	2.09	19.36	1.75
LLaMA-VID (Li, Wang, and Jia 2023)	34.14	2.21	25.02	1.99	27.18	2.00	22.16	1.84
Chat-UniVi (Jin et al. 2023)	35.61	2.23	25.89	2.01	33.23	2.13	20.59	1.79
Video-LLaVA (Lin et al. 2023)	39.46	2.37	30.78	2.15	32.95	2.18	24.31	1.90
VLM-RLAIF† (Ahn et al. 2024)	51.16	2.55	41.44	2.30	46.52	2.41	29.78	1.94
PLLaVA† (Xu et al. 2024)	48.92	2.53	38.26	2.28	43.83	2.40	30.92	2.07
LLaVA-NeXT-DPO† (Zhang et al. 2024b)	65.08	2.82	59.12	2.65	60.80	<u>2.70</u>	40.14	2.24
LLaVA-Hound-DPO (Zhang et al. 2024a)	<u>73.64</u>	<u>3.12</u>	<u>68.29</u>	<u>2.98</u>	<u>74.00</u>	**3.12**	<u>48.89</u>	<u>2.53</u>
ISR-DPO (π_{θ^1})	74.33	3.12	68.18	2.96	73.57	3.10	48.91	2.52
ISR-DPO (π_{θ^5})	79.63	3.19	74.07	3.05	77.52	3.12	53.13	2.57
ISR-DPO (π_{θ^9})	**80.36**	**3.20**	**75.42**	**3.05**	**78.58**	**3.12**	**54.66**	**2.59**

Table 2: **Quantitative comparison between different VLMMs on *out-domain* video question answering with detailed captions as supporting evidence proposed in Zhang et al. (2024a).** Our final iterated model of ISR-DPO (π_{θ^9}) consistently outperforms all other models in both accuracy and score across these benchmarks, demonstrating superior performance in out-domain video question answering tasks. The best results are **bold** and the second-best results are <u>underlined</u>. †: reproduced by the authors' implementation. All results except † are directly sourced from Zhang et al. (2024a).

lengths in ISR-DPO. We posit that dual conditioning on video content and visual context during preference selection enables the VLMM to select preferences based on video information rather than length bias. This results in a lower chosen-to-rejected preference ratio and shorter, more concise responses from the VLMM, as illustrated in Fig. 5.

Moreover, we compare the $9th$ iteration model's responses between self-rewarding and ISR-DPO to validate the effectiveness of visual context, as in Yuan et al. (2024). In particular, we use GPT-4 as the evaluator by selecting the response closest to the ground truth, assessing win-rates. Figure 6 shows the win-rate between self-rewarding and ISR-DPO across all benchmarks, demonstrating the effectiveness of ISR-DPO. Notably, despite generating more concise responses (Fig. 5), ISR-DPO consistently achieved higher winning rates across all benchmarks. This provides evidence of ISR-DPO's effectiveness at conveying more relevant and accurate information within concise responses, mitigating verbosity hallucinations.

Effect of visual context on human alignment. To evaluate the impact of visual context on judgment quality, we assess the correspondence between AI models' preferences and those of human annotators, following Lee et al. (2023). As shown in Tab. 4, ISR-DPO demonstrates a higher human alignment accuracy (75.0 %) compared to self-rewarding (59.0 %), suggesting that the incorporation of visual context enhances the model's ability to make human-like assessments.

Methods	MSVD-QA		MSRVTT-QA		TGIF-QA	
	Acc.	Score	Acc.	Score	Acc.	Score
Video-ChatGPT (Maaz et al. 2024)	68.6	3.8	58.9	3.4	47.8	3.2
Chat-UniVi (Jin et al. 2023)	70.0	3.9	53.1	3.1	46.1	3.1
VideoChat2 (Li et al. 2024)	70.0	3.9	54.1	3.3	-	-
Video-LLaVA (Lin et al. 2023)	71.8	3.9	59.0	3.4	48.4	3.2
LLaMA-VID (Li, Wang, and Jia 2023)	72.6	3.9	58.7	3.4	49.2	3.3
PLLaVA† (Xu et al. 2024)	78.8	4.0	65.6	3.4	57.9	**3.5**
LLaVA-NeXT-DPO† (Zhang et al. 2024b)	78.6	4.0	63.4	3.1	58.2	<u>3.4</u>
VLM-RLAIF† (Ahn et al. 2024)	<u>81.0</u>	<u>4.2</u>	69.2	<u>3.7</u>	<u>62.3</u>	**3.5**
LLaVA-Hound-DPO (Zhang et al. 2024a)	80.7	4.1	<u>70.2</u>	<u>3.7</u>	61.4	**3.5**
ISR-DPO (π_{θ^1})	80.1	4.1	69.8	3.6	61.0	3.4
ISR-DPO (π_{θ^5})	84.8	4.3	76.0	3.8	66.8	3.5
ISR-DPO (π_{θ^9})	**85.8**	**4.3**	**78.7**	**3.9**	**67.8**	**3.5**

Table 3: **Comparison of different VLMMs on *out-domain* video question answering benchmark (Maaz et al. 2024).** ISR-DPO (π_{θ^9}) outperforms previous work across three video question answering datasets. Best results in **bold**, second-best <u>underlined</u>. †: reproduced with the authors' implementation. eOther results are directly sourced from Zhang et al. (2024a).

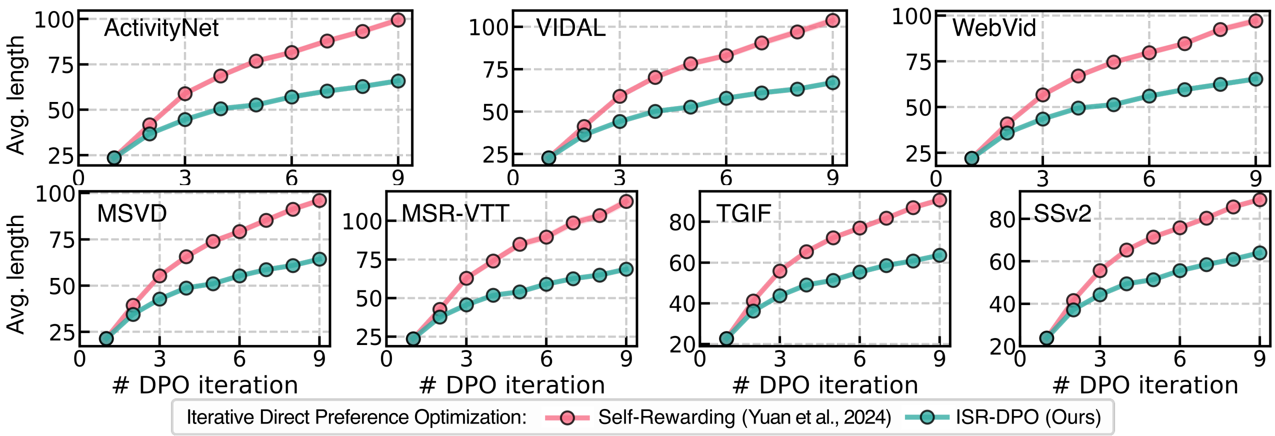

Figure 5: **Average (Avg.) response word length between self-rewarding and ISR-DPO on various video question answering benchmarks.** ISR-DPO yields compact and concise responses at the same iteration compared to self-rewarding.

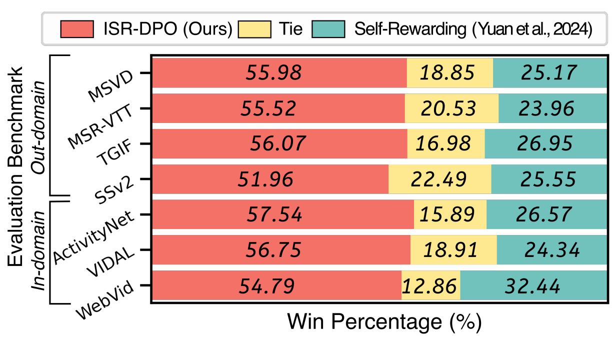

Figure 6: **Head-to-head performance comparison at 9th iteration.** ISR-DPO consistently outperforms the self-rewarding across benchmarks.

Task	Human Alignment Accuracy (%)	
	Self-rewarding	ISR-DPO
Preference selection	59.0	75.0

Table 4: **Human annotator alignment accuracy for preference selection.** We measure human alignment accuracy to evaluate the amount of correlation between human and aligned models, *i.e.*, self-rewarding *vs*. ISR-DPO.

Various design choices for visual context. We examine various design choices for visual context in Tab. 5: (1) without context ('N/A'), (2) Fixed context from the first iteration ('Fixed'), (3) New context at each iteration ('Renew') and (4) Self-retrospective context ('Self-retro.'). The 'Self-retro.' consistently performs the best, leveraging and refining previous context while adding details with improved video understanding (Fig. 7). Interestingly, 'Fixed' outperforms 'Renew' in most benchmarks, except for MSRVTT. For SSv2 and WebVid, 'Renew' even performs worse than 'N/A'. We hypothesize that 'Renew' may introduce incon-

Context Design	Out-of-domain Video QA Benchmark								In-domain Video QA Benchmark					
	MSVD		MSRVTT		TGIF		SSV2		ActivityNet		VIDAL		WebVid	
	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
N/A	78.73	3.14	73.42	3.00	77.10	3.09	54.34	2.56	81.96	3.23	76.71	3.09	87.24	3.39
Fixed	79.17	3.15	74.35	3.02	77.88	3.09	54.29	2.57	82.25	3.24	77.90	3.12	87.49	3.39
Renew	79.49	3.19	74.04	3.04	77.63	3.12	53.03	2.56	82.03	3.26	77.73	3.12	86.68	3.38
Retrospective	**80.36**	**3.20**	**75.42**	**3.05**	**78.58**	**3.12**	**54.66**	**2.59**	**82.99**	**3.26**	**79.00**	**3.13**	**88.11**	**3.40**

Table 5: **Quantitative comparison of various designs for generating visual context.** 'N/A' indicates no use of context, 'Fixed' uses context generated in the first iteration for all subsequent iterations, 'Renew' generates new context each iteration, and 'Retrospective.' employs a self-retrospective context.

Figure 7: **Visualization of predicted context over iteration.** Generated context becomes increasingly well-grounded over iteration. Red indicates irrelevant responses, while blue indicates accurate, visually-grounded responses.

Figure 8: **Qualitative comparison of self-rewarding *vs.* ISR-DPO.** The figure contrasts descriptions generated without visual context, *i.e.*, self-rewarding (upper), against those with visual context, *i.e.*, ISR-DPO (bottom), at the $9th$ iteration. Visual context results in more accurate, concise, and relevant descriptions. Red indicates irrelevant responses, while blue indicates well-grounded responses.

sistent focus in the video across iterations, potentially causing attention to irrelevant details. These findings suggest that a methodical approach to context renewal, such as our 'Self-retro.', is crucial for maintaining focus on relevant content, thereby improving proper preference modeling.

4.4 Qualitative Analysis

Enhanced visual context over iteration. To show the improving nature of self-retrospective context, we visualize the generated context as shown in Fig. 7. As training iterations progress, the context adds more and more detailed visual information about the video, such as specific species of goldfish. This improved context aids the overall understanding of the video content to improve preference selection process.

Comparison of self-rewarding *vs.* ISR-DPO. Figure 8 compares the responses of self-rewarding, *i.e.*, ISR-DPO w/o visual context, and ISR-DPO for $9th$ iterated models. The self-rewarding tends to produce longer responses, but as sentences progress, they become less relevant to the question and visual content. Also, it fails to recognize the athlete's jumping motion accurately. In contrast, ISR-DPO generates more concise and accurate responses that are well-grounded in the video content.

5 Conclusion

We present ISR-DPO, a novel iterative direct preference optimization for VLMMs that enhances the instruction-following ability for videos. In particular, we propose self-retrospective preference modeling to improve VLMM's capability to judge visually grounded preferences. By doing so, ISR-DPO mitigates the model's problematic inclination for visually ungrounded verbosity in judging preferred response, leading to more concise and visually grounded responses. Empirical evaluations across various video question answering benchmarks demonstrate ISR-DPO's superior performance compared to the state of the art VLMMs.

6 Acknowledgment

This work was partly supported by CARAI grant funded by DAPA and ADD (UD230017TD) and the IITP grants (No.RS-2022-II220077, No.RS-2022-II220113, No.RS-2022-II220959, No.RS-2022-II220871, No.RS-2021-II211343 (SNU AI), No.RS-2021-II212068 (AI Innov. Hub)) funded by the Korea government(MSIT).

References

Ahn, D.; Choi, Y.; Yu, Y.; Kang, D.; and Choi, J. 2024. Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback. In *ACL*.

Anderson, R. C. 1984. Schema Theory: An Introduction. In Anderson, R. C.; Osborn, J.; and Tierney, R. J., eds., *Learning to Read in American Schools: Basal Readers and Content Texts*, 243–258. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bransford, J. D.; and Johnson, M. K. 1972. Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6): 717–726.

Burns, C.; Izmailov, P.; Kirchner, J. H.; Baker, B.; Gao, L.; Aschenbrenner, L.; Chen, Y.; Ecoffet, A.; Joglekar, M.; Leike, J.; Sutskever, I.; and Wu, J. 2023. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. arXiv:2312.09390.

Chen, Z.; Deng, Y.; Yuan, H.; Ji, K.; and Gu, Q. 2024. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. In *ICML*.

Dweck, C. S. 2006. *Mindset: The new psychology of success*. New York: Random House. ISBN 978-1400062751.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jin, P.; Takanobu, R.; Zhang, C.; Cao, X.; and Yuan, L. 2023. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv preprint arXiv:2311.08046*.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2): 163–182.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; and Prakash, S. 2023. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:2309.00267*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; Wang, L.; and Qiao, Y. 2024. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *arXiv*.

Li, Y.; Wang, C.; and Jia, J. 2023. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. *arXiv preprint arXiv:2311.17043*.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*.

Liu, R.; Li, C.; Tang, H.; Ge, Y.; Shan, Y.; and Li, G. 2023. ST-LLM: Large Language Models Are Effective Temporal Learners. *https://arxiv.org/abs/2404.00308*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pang, R. Y.; Yuan, W.; Cho, K.; He, H.; Sukhbaatar, S.; and Weston, J. 2024. Iterative Reasoning Preference Optimization. arXiv:2404.19733.

Park, R.; Rafailov, R.; Ermon, S.; and Finn, C. 2024. Disentangling Length from Quality in Direct Preference Optimization. arXiv:2403.19159.

Prasann Singhal, J. X., Tanya Goyal; and Durrett, G. 2023. A Long Way to Go: Investigating Length Correlations in RLHF. *arXiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 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*.

Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Schapire, R. E. 1990. The strength of weak learnability. *Machine Learning*, 5(2): 197–227.

Simon, H. A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6): 467–482.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *ACM Multimedia*.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. arXiv:2404.16994.

Yuan, W.; Pang, R. Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-Rewarding Language Models. arXiv:2401.10020.

Zhang, R.; Gui, L.; Sun, Z.; Feng, Y.; Xu, K.; Zhang, Y.; Fu, D.; Li, C.; Hauptmann, A.; Bisk, Y.; and Yang, Y. 2024a. Direct Preference Optimization of Video Large Multimodal Models from Language Model Reward. arXiv:2404.01258.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024b. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024. Calibrated Self-Rewarding Vision Language Models. arXiv:2405.14622.

7 Additional Input Prompts for Preference Dataset Generation

In the process of generating our preference dataset, we employ specific additional input prompts for each stage. Figure 9 illustrates three types of input prompts used in this process: 1) response generation, 2) self-retrospective context generation, and 3) preference judgment. The 'Prompt (response)' defines the guideline for VLMM's responses and is used consistently throughout all stages of data generation. The 'Prompt (context)' demonstrates the prompt used to generate a context based on the previous context. Lastly, the 'Prompt (judge)' presents the prompt used for preference judgment using the current Video Large Multimodal Model (VLMM).

8 Details on Head-to-Head Comparison with GPT-4 Evaluator

We evaluate the generated response quality of the ISR-DPO compared to self-rewarding (Yuan et al. 2024) through a head-to-head comparison (Yuan et al. 2024). Specifically, we prompted GPT-4 to determine which of the two responses is superior across in-domain and out-of-domain video question answering benchmarks. The evaluation focused on two key aspects: 1) the relevance of model's answer to the provided instruction, and 2) the accuracy of the response in relation to the ground-truth answer. We visualize a detailed prompt in Fig. 10.

9 Details on Human Evaluation for Human Preference Alignment

We conduct a human evaluation to measure how well the AI-generated preferences align with human preferences, following the approach of Lee et al. (2023). We randomly sample 100 questions from the validation set of video question-answering dataset (Xu et al. 2017). We then recruit 15 annotators per question through the Amazon Mechanical Turk platform. Annotators are presented with a video, an instruction, and two versions of responses generated from our ISR-DPO. Specific instructions and examples of the questions given to the annotators can be found in Fig. 11.

10 More Qualitative Results

In Fig. 12, we present additional examples comparing responses generated by self-rewarding and our ISR-DPO. Well-grounded phrases are highlighted in blue, while misaligned or irrelevant phrases are marked in red. Compared to self-rewarding, our approach reduces the occurrence of misaligned and overly verbose sentences. For instance, in the beach soccer example, our method accurately identifies the team colors as blue and orange without unnecessary elaboration. These examples demonstrate how our ISR-DPO reduces verbosity hallucination, generating more concise and relevant responses.

11 Performance Over Training Iterations

In Fig. 13, we demonstrate the effectiveness of ISR-DPO across training iterations using various video question answering benchmarks for evaluation. Overall, the performance improves as we increase the number of training iterations, with the exception of the MSR-VTT dataset at the 7th iteration. However, we can observe that the performance recovers and improves again in subsequent training iterations up to the 9th iteration.

Prompt (response)

```
A chat between a curious human and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the human's questions.
```

Prompt (context)

```
Given the provided video and the previously predicted description of the video, your task is to generate an
enhanced description of the video clip. The generated description should provide a comprehensive
understanding of the video's content while forming a coherent story.
Note that the previous description might include irrelevant or inappropriate words. Thus, you don't have to
include all the contents in the previous description. Focus on generating new description with improved
accuracy and detail, while it concise as possible and comprehensive.
--------------------------------------------------
Consider the following aspects while generating the description:
(1) Unwavering Accuracy: The AI must consistently provide accurate and verifiable information in its
responses.
(2) Consistent Calibration: The AI should ensure that its confidence assertions align with its actual
accuracy, reflecting genuine knowledge estimation.
(3) Transparent Uncertainty Expression: When uncertain, the AI should transparently express its level of
uncertainty, ensuring users are never misled by vague or overly confident statements.
(4) Honest Self-representation: The AI must be forthright about its own capabilities and should not feign
expertise or knowledge it doesn't possess.
(5) Additional guidelines:
  1) Clarify the predicted answer by avoiding issues such as repetition, unclear descriptions, or any
grammatical errors that could hinder understanding.
  2) Address inconsistencies to ensure accuracy and alignment with the video's content.
  3) Generate a description concisely.
--------------------------------------------------
Previous description:
<PREVIOUS CONTEXT>
--------------------------------------------------
Now, generate the improved description below.
Improved description:
```

Prompt (judge)

```
Considering given video, question and description about the video, select the most preferred (least
wrong) answer for the question.

Inputs:
 1. Question, the question queried to the AI system
 2. Answer1, the first answer prediction from the AI system
 3. Answer2, the second answer prediction from the AI system
 4. Video description, the overall description about the given video in the form of a paragraph

Based on the provided definition, please select the preferred answer (Answer1 or Answer2) for the
given instruction (Question) and provide a concise explanation for choosing it as the correct one.
When generating the explanation, please provide valid justifications without simply mentioning the
answer itself.

Your answer should be formatted as:
Choice:
Explanation:

Now provide your answer in this situation:
Question: <QUESTION>
Answer1: <PREDICTION A>
Answer2: <PREDICTION B>
Description: <RETRO CONTEXT>
 When generating output, you should consider the visual situation provided and include either
'Answer1' or 'Answer2' in your generated output.
```

Figure 9: **Various input prompts for constructing preference dataset.** This shows various input prompts: the upper part for generating two responses, the center part for context generation based on previous context, and the bottom part for preference judgment using the VLMM from the latest iteration.

Given the following inputs:

1. **Question Related to the Caption**: **{Question}**
2. **Ground Truth Answer**: **{Answer}**
3. **Response1**: **{Prediction1}**
4. **Response2**: **{Prediction2}**

Your task is to evaluate which of two model's outputs is better, based on the ground truth answer and the question. Consider the following criteria for evaluation:
You must choose either Response1 or Response2 as better than the other.

- **Relevance**: Does the predicted answer directly address the question posed? The response should not contain uncorrelated sentences with respect to the question. For example, if the question asks about the man in the video, only describe about the object and not the background, atmosphere, etc.
- **Accuracy**: Compare the predicted answer to the ground truth answer. Does the prediction accurately reflect the information given in the ground truth answer without introducing factual inaccuracies?
Note: For answering Choice, you should respond with either Response1, Response2 or Tie. DO NOT PROVIDE ANY EXPLANATION. Except when unavoidable, please avoid using Tie and choose either Response1 or Response2.

Output Format:
Choice: <choice of better response: 1 for Response1, 2 for Response2 and 3 for Tie>

Figure 10: **Evaluation criteria provided to GPT-4.** To compare the generated responses of self-rewarding and ISR-DPO, we prompted GPT-4 to choose better response regrading two criteria: Relevance and precision.

Task:
As an annotator, your task is to evaluate two AI-generated responses to a question related to a specific video and its caption. You must choose the response in relation to the question asked and the information available in the video, that better fulfills the criteria described below. Use the information available in the visualized video frames and the provided ground-truth answer to assess the response.

Criteria for Evaluation:
1. **Accuracy**: Assess whether the response accurately answers the question, adhering closely to the content and context provided by the video. The ideal response should precisely reflect the intent and factual content of the ground-truth answer. Avoid selecting responses that introduce information not present in the video and ground-truth. The response should not speculate, assume, or create unfounded details, regardless of their plausibility. Conciseness

2. **Relevance**: The response should directly address the question without including unnecessary sentences or digressions. Check that the response maintains relevance to the question throughout. Responses should not only start relevantly but should also maintain this focus until the end. - e.g.) If the question concerns "What is the man's attire?", penalize responses that wander into discussing irrelevant details such as the overall atmosphere of the video, weather conditions, etc., unless these elements are quite relevant to the question.

3. **Sentence Structure**: The response should be well-structured and easy to understand. It should use clear language and logical sequencing of ideas. Avoid responses that are ambiguous or poorly constructed, as these can lead to misinterpretation of the information.

Procedure (FYI):
- Read the question provided.
- Read the ground-truth answer and understand the specific focus of the question.
- Review both AI-generated responses.
- If there are additional information that is hard to fact-check with ground-truth caption, check the visualized video frames to decide wheter the information is correct.
- Review both AI-generated responses.
- Select the response that best meets the criteria overall.

Notes:
In cases where both responses seem equally valid, consider subtle differences in tone, the precision of language, and the logical flow of ideas. Remember that the ultimate goal is to select the response that best serves a user seeking information about the video and question posed, reflecting both the content and spirit of the source material. Your careful analysis and selections are crucial in helping improve the accuracy and usability of AI-generated content. Thank you for your diligent efforts in this task.

*

Question: What is the location and date of the "NIGHT OF THE PROMS" event mentioned in the video?

Ground-truth Answer: According to the video caption, the "NIGHT OF THE PROMS" event will occur for the first time in Poland on the 22nd of March in the city of Łódź.

Response 1: The "NIGHT OF THE PROMS" event mentioned in the video is taking place on March 22, 2019, at the "Lodz" location, as indicated by the text overlays in the video. The setting suggests that this is a live stage event, and the location and date are provided to contextualize the setting and timing of the event.

Response 2: The event mentioned in the video is the "NIGHT OF THE PROMS," which is a televised singing competition or concert. The text overlays in the video indicate that the event took place in Poland, with specific dates mentioned for different stages or rounds of the competition. The text overlays also provide the location and date, suggesting that the event was a televised singing competition or concert that took place in Poland, with various stages or rounds indicated by the dates. The event was broadcasted or took place in 2016.

https://drive.google.com/uc?export=view&id=1FGN1NXD5i8hsuocCoMFaVj82szfDEgil

○ Response 1

○ Response 2

Figure 11: **Evaluation criteria provided to Amazon Mechanical Turk annotators.** We carefully instructed the annotators to penalize the outputs that include unaligned contents with the provided video, or the answer that contains overly verbose sentences that deviates from the question's purposes.

Question: What is the attire of the teams playing in the beach soccer match?

Ground-Truth

The teams are wearing two distinct kits, with one team in blue and the other in orange.

Self-Rewarding (Yuan et al., 2024)

The teams playing in the beach soccer match are dressed in distinctive attire that contrasts with each other, making it easy to distinguish between the two sides. The players are wearing team uniforms that are typically shorts and jerseys, suitable for the beach environment and the physical nature of the sport. The colors of the uniforms are vibrant and likely represent their respective teams, with one team wearing lighter colors and the other in darker hues. The attire is appropriate for the sport and the setting, allowing for ease of movement and comfort during the match.

ISR-DPO (Ours)

The teams playing in the beach soccer match are dressed in distinctive team colors, with one team wearing **blue** and the other in **orange**. Their attire is appropriate for beach sports, consisting of shorts and T-shirts or jerseys, which allow for agility and comfort during the game. The players' uniforms are vibrant and contrasting, making it easy to distinguish between the two teams.

Question: What is the weight of the discus being thrown?

Ground-Truth

The weight of the discus being thrown is 1.5kg according to visible text annotations.

Self-Rewarding (Yuan et al., 2024)

The weight of the discus being thrown is prominently displayed in the video, with text overlays indicating the weight of the discus at various points during the throw. The discus is thrown with a specific weight in mind, which is crucial for the performance and technique of the athlete. The weight of the discus being thrown is 5 kg, 5.5 kg, and 6 kg, as shown in the text overlays in the video. These weights are significant as they provide context for the athlete's strength and the technical requirements of the throw.

ISR-DPO (Ours)

The weight of the discus being thrown is indicated by text overlays in the video, which specify the weight of the discus as "1.5kg". These details provide viewers with information about the weight of the discus being used in the throw.

Question: What kind of chips are arranged on the white round plate in the video?

Ground-Truth

Triangular yellow tortilla chips are arranged in a circular pattern on the white round plate in the video.

Self-Rewarding (Yuan et al., 2024)

The chips arranged on the white round plate in the video are yellow corn chips, which are commonly known for their crisp texture and mildly salty flavor. They are typically made from cornmeal and appear to be of a standard size and shape, commonly found in snack bags or served at casual dining establishments. The chips are arranged in a way that suggests they are ready to be dipped or eaten with the cheese sauce.

ISR-DPO (Ours)

The chips arranged on the white round plate in the video appear to be tortilla chips, which are commonly used as a dipping medium for such sauces. They have a triangular shape and a golden-yellow hue, indicative of being fried and seasoned. These chips are typically made from cornmeal and are a popular accompaniment to dips and sauces.

Question: What kind of protective gear is the individual wearing in the video?

Ground-Truth

The individual in the video is wearing a welding helmet and gloves as protective gear.

Self-Rewarding (Yuan et al., 2024)

The individual in the video is wearing protective gear appropriate for welding, including a helmet with a darkened visor to shield their face from the intense light and sparks, gloves to protect their hands, and a long-sleeved shirt to safeguard their skin from burns. The gear is designed to protect the welder from the hazards of the welding process. The protective clothing is primarily dark in color, likely chosen for its durability and ability to absorb heat. The helmet is also equipped with a face shield to ensure the welder's face is fully protected.

ISR-DPO (Ours)

The individual in the video is wearing protective gear, including a welding helmet and gloves, which are essential for safety during such operations. The protective gear suggests that the individual is taking necessary precautions to shield themselves from the intense light and potential hazards of the welding process. The gear is typical for welding tasks, ensuring the individual's safety while performing the work.

Figure 12: **More qualitative example of prediction from self-rewarding *vs.* ISR-DPO.** We compare responses generated at the 9*th* iteration for both models. Integrating visual context leads to more accurate, concise, and relevant descriptions that align more closely with the ground-truth answer. Red indicates irrelevant or wrong responses, while blue indicates well-grounded responses.

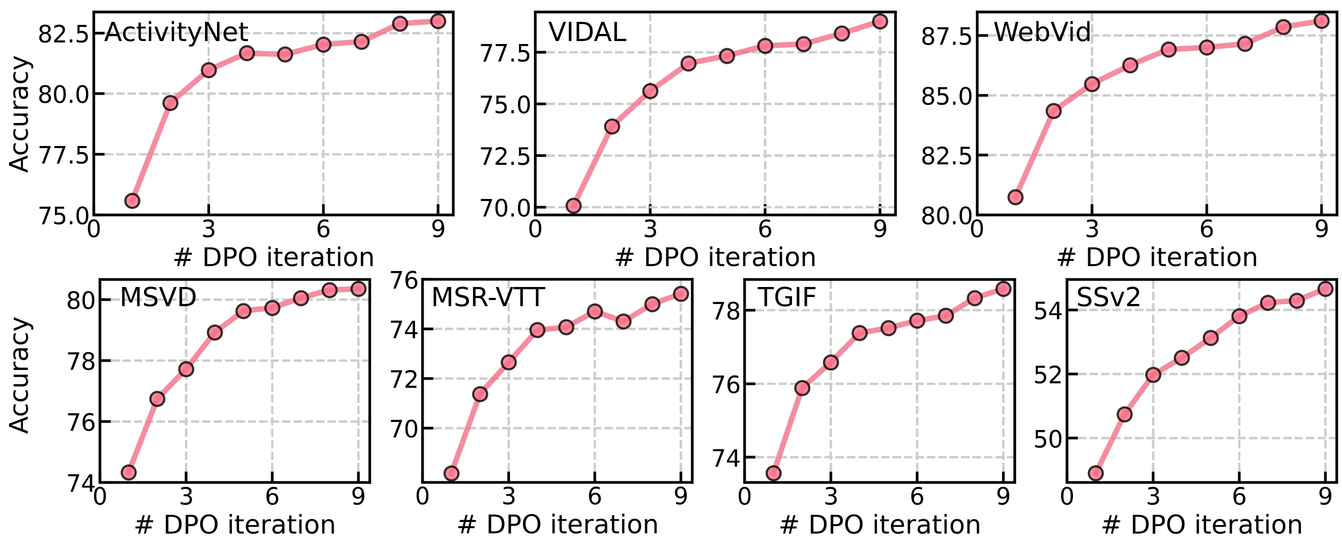

Figure 13: **Accuracy of ISR-DPO over iterations on video question answering benchmarks.** Overall, our ISR-DPO consistently improves its performance over DPO iteration. In-domain datasets: Activity-Net, VIDAL and WebVid, Out-domain datasets: MSVD, MSR-VTT, TGIF and SSv2 used in Zhang et al. (2024a).