Self-alignment of Large Video Language Models with Refined Regularized Preference Optimization

Pritam Sarkar
Queen's University, Canada and Vector Institute
pritam.sarkar@queensu.ca

Ali Etemad
Queen's University, Canada
ali.etemad@queensu.ca

Website **Code** **Models**

Abstract

Despite recent advances in Large Video Language Models (LVLMs), they still struggle with fine-grained temporal understanding, hallucinate, and often make simple mistakes on even simple video question-answering tasks, all of which pose significant challenges to their safe and reliable deployment in real-world applications. To address these limitations, we propose a self-alignment framework that enables LVLMs to learn from their own errors. Our proposed framework first obtains a training set of preferred and non-preferred response pairs, where non-preferred responses are generated by incorporating common error patterns that often occur due to inadequate spatio-temporal understanding, spurious correlations between co-occurring concepts, and over-reliance on linguistic cues while neglecting the vision modality, among others. To facilitate self-alignment of LVLMs with the constructed preferred and non-preferred response pairs, we introduce Refined Regularized Preference Optimization (RRPO), a novel preference optimization method that utilizes sub-sequence-level refined rewards and token-wise KL regularization to address the limitations of Direct Preference Optimization (DPO). We demonstrate that RRPO achieves more precise alignment and more stable training compared to DPO. Our experiments and analysis validate the effectiveness of our approach across diverse video tasks, including video hallucination, short- and long-video understanding, and fine-grained temporal reasoning.

1 Introduction

Despite recent progress in Large Video Language Models (LVLMs) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15], these models continue to face limitations in fine-grained temporal understanding [16, 17, 18], demonstrate a propensity for hallucination [19, 20], struggle with long-video understanding [21, 22, 23, 24, 25], and frequently make naive mistakes in short-video question-answering tasks [16, 17, 18]. Please see Figure 1 for a number of examples. These shortcomings severely limit their safe and reliable deployment in real-world applications. The underlying causes

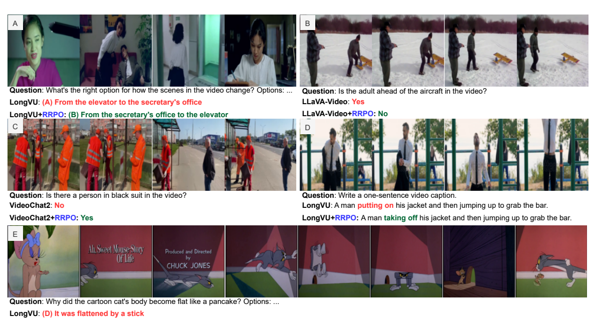


Figure 1: Examples of failure cases in simple video understanding tasks by state-of-the-art LVLMs, and improvements observed after self-alignment with RRPO.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

of these limitations are multifaceted, encompassing factors such as inadequate spatial and temporal understanding [16, 17, 18], vision-language representational misalignments [26, 22], challenges in processing long visual sequences due to context length constraints [21, 22, 23], spurious correlations between co-occurring concepts [27, 28], and an over-reliance on linguistic cues while neglecting the visual information [16, 17, 18, 29].

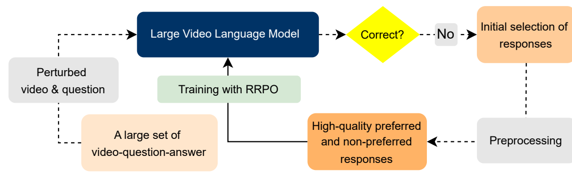


Figure 2: An overview of our self-alignment framework.

To address these shortcomings and enhance video understanding in LVLMs, we design a self-alignment [30] framework that enables LVLMs to learn from their own errors. Specifically, we begin by sampling video-question pairs from an open-source benchmark. We then apply spatio-temporal perturbations to the video content to mimic common errors that often arise from over-reliance on linguistic cues, spurious correlations between co-occurring concepts, and insufficient spatio-temporal understanding. Using the perturbed video and the corresponding question, we perform inference with the target LVLM. If the model's response is incorrect, we construct a self-alignment pair by treating the incorrect response as a non-preferred sample and the correct response as the preferred one, which is kept for self-alignment training. Responses that are already correct are discarded, as they offer limited self-improvement potential. Subsequently, we optimize a loss function to prioritize the preferred response over the non-preferred one. Notably, our data generation pipeline is free from human annotation and can be easily scaled. A high-level overview of our self-alignment framework is depicted in Figure 2.

To effectively train LVLMs using the generated preferred and non-preferred response pairs, we introduce ***Refined Regularized Preference Optimization (RRPO)***, an approach designed to address limitations of Direct Preference Optimization (DPO) [31]. RRPO is designed to address the key drawbacks of DPO. RRPO provides a fine-grained sub-sequence-level reward to explicitly penalize the tokens containing the key concepts that are different between the preferred and non-preferred response pairs. This is contrary to DPO's response-level reward which penalizes all tokens within non-preferred responses, thus lacking precision and often proving unsuitable for fine-grained alignment. It should be noted that our fine-grained reward further benefits from computing a smaller gradient during optimization and is therefore less prone to diverge away from its initial state unlike DPO's response-level feedback along with weak regularization which can cause significant divergence from the base model, leading to suboptimal performance. However, sub-sequence-level rewards may incentivize the model to exploit shortcuts, such as outputting correct concepts without proper context or complete responses. To mitigate this, we also minimize token-wise KL-divergence [27] using a reference model on the preferred responses. This ensures that the LVLM maintains a tight bound on its likelihood across the entirety of the preferred response while reducing the likelihood on the non-preferred concepts.

Through empirical and theoretical analysis, we demonstrate that RRPO exhibits greater stability and smoother convergence during optimization compared to DPO. We validate our method on three popular LVLMs specialized for video understanding, VideoChat2, LLaVA-Video, and longVU, covering a wide range of architectures, LLMs, vision encoders, and training setups. Our in-depth evaluation demonstrates that our proposed self-alignment reduces hallucinations and improves performance in fine-grained temporal understanding, as well as in both short- and long-video understanding tasks.

In summary, our contributions are as follows:

- We design a *self-alignment* framework to facilitate self-improvement of LVLMs based on their own errors. We introduce **RRPO**, a preference optimization method that addresses the limitations of DPO by utilizing sub-sequence-level refined rewards and token-wise strong KL regularizer, resulting in more precise alignment and stable training.
- Our rigorous evaluation demonstrates the effectiveness of our proposed method across diverse video tasks, including video hallucination, short and long video understanding, and fine-grained temporal reasoning, among others. Moreover, our experimental and theoretical analyses highlight the superiority of RRPO over DPO in aligning LVLMs. We make our code, data, and model weights public to enable fast and accurate reproducibility.

2 Preference Responses for Self-alignment

As the first step in our framework, we construct a training dataset $\mathcal{D}$ comprising both preferred responses and responses containing incorrect concepts in order to align the LVLM π_θ to favor correct concepts over incorrect ones. We begin by utilizing a large and diverse publicly available video instruction tuning dataset, containing triplets of video x_v, question x_q, and their human-preferred answers y^+. To generate non-preferred responses y^-, we obtain perturbed videos $\hat{x}_v = f_p(x_v)$, where f_p is a perturbation function which masks a large portion of frames and applies temporal shuffling, compromising spatio-temporal consistency. Our intention from this step is to provoke the LVLM to generate responses primarily based on language cues or their parametric knowledge. These perturbed videos $\hat{x}_v$ are fed to π_θ as inputs to generate responses with potential erroneous concepts $y^- = \pi_\theta(x_q, \hat{x}_v)$. An example is illustrated in Figure 3. We then verify the correctness of y^-, retaining incorrect responses and discarding correct ones. Next, We employ an LLM to meticulously compare y^- against y^+ and identify key incorrect concepts within y^-, ensuring that for each incorrect concept $y_i^- \in y^-$ there exists a corresponding correct concept $y_i^+ \in y^+$. In the context of our work, a concept can be actions, objects, attributes, relations, and other key elements in the response that contribute to the semantic understanding of the video. Furthermore, for lengthy responses, we maintain structural similarity between the incorrect and correct versions by rewriting the correct response while incorporating the incorrect concepts. Finally, we apply deduplication based on these correct-incorrect concept pairs, constructing a high-quality training dataset. Examples are provided in Figure 4.

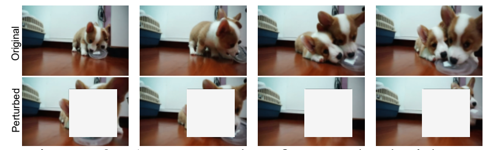


Figure 3: An example of perturbed video.

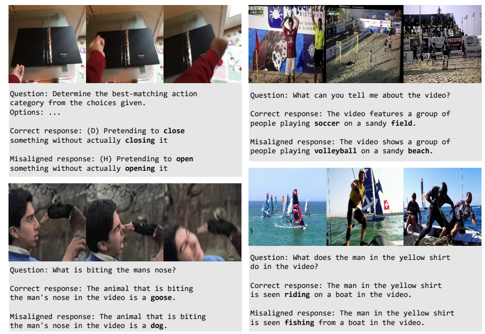


Figure 4: A few training samples. We bold the correct and incorrect concepts.

3 RRPO

Preliminaries. Given an input $x = \{x_q, x_v\}$ with a pair of responses $\{y^+, y^-\}$, where $y^+ \succ y^-|x$, DPO [31] can be employed to align π_θ to favor y^+ over y^-. This is achieved by maximizing the reward margin between π_θ and a reference model π_{ref}, using the following training objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\Big[\log\sigma\big(r_\theta(x, y^+) - r_\theta(x, y^-)\big)\Big], \tag{1}$$

where reward $r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ and β controls the deviation from π_{ref}. Note that $r_\theta(x, y)$ is calculated at a response-level by penalizing all the tokens in y, despite the fact that the difference between y^+ and y^- might only be limited to a *few* key tokens. Such response-level reward can be considered coarse-grained reward modeling and is not suitable for fine-grained alignment. Moreover, as the reward is calculated by penalizing all the tokens in the response, the gradient for $\mathcal{L}_{\text{DPO}}$ tends to be very large for long responses, and accordingly π_θ can diverge to an undesired state thus losing its out-of-the-box capabilities [27, 32, 33].

Refined Regularized Preference Optimization. Our goal is to design a method that can provide a fine-grained reward to penalize individual sub-sequences consisting of the tokens belong to the key differing concepts between y^+ and y^-. We refer to this as refined reward modeling. Let y be expressed as a sequence of tokens $T = \{t_1, t_2, \ldots, t_{|y|}\}$. Here, the i-th sub-sequence y_i can be expressed as $T[s_i : e_i]$, where s_i and e_i are the start and end indices of y_i with $1 \leq s_i \leq e_i \leq |y|$. Therefore, the reward for y_i can be computed as:

$$r_\theta(x, y_i) = \beta \log \frac{\prod_{j=s_i}^{e_i} \pi_\theta(t_j|x, t_{<j})}{\prod_{j=s_i}^{e_i} \pi_{\text{ref}}(t_j|x, t_{<j})}. \tag{2}$$

Assuming there exists N such sub-sequences, we can train π_θ to maximize the total reward margin

$$u = \sum_{i=1}^{N} u_i = \sum_{i=1}^{N} \big(r_\theta(x, y_i^+) - r_\theta(x, y_i^-) \big). \tag{3}$$

Subsequently, we replace the reward formulation in Equation 1 with our refined reward modeling as:

$$\mathcal{L}_{\text{RRPO}}^{(\text{rank})}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \Big[\log \sigma(u) \Big]. \tag{4}$$

However, the sparsity of the rewards used to calculate $\mathcal{L}_{\text{RRPO}}^{(\text{rank})}$ may allow π_θ to exploit shortcuts and effectively game the reward function by merely outputting key concepts without appropriate context or generating complete responses. This *reward hacking* may occur since π_θ is incentivized to maximize reward margin based on the differing sub-sequences; the model can learn to produce short, incomplete responses that contain the correct key concepts, even if those responses lack overall coherence, fluency, or completeness. To mitigate this reward hacking, we introduce a regularizer between π_θ and π_{ref}, based on token-wise KL (TKL) divergence [27, 34], as follows:

$$\mathbb{D}_{\text{TKL}}\big(x, y; \pi_{\text{ref}} \,\|\, \pi_\theta\big) = \sum_{t=1}^{|y|} \mathbb{D}_{\text{KL}}\big(\pi_{\text{ref}}(\cdot \mid [x, y_{<t}]) \,\|\, \pi_\theta(\cdot \mid [x, y_{<t}])\big). \tag{5}$$

Accordingly, to ensure π_θ retains its original likelihood across the entirety of y^+ while reducing the likelihood of non-preferred concepts, we optimize $\mathbb{D}_{\text{TKL}}$ between π_θ and π_{ref} over y^+. We then modify Equation 4 and define the final RRPO loss as:

$$\mathcal{L}_{\text{RRPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \Big[\log \sigma(u) + \alpha \cdot \mathbb{D}_{\text{TKL}}\big(x, y^+\big) \Big], \tag{6}$$

where α controls the divergence between π_θ and π_{ref}, and π_{ref} is kept fixed.

How is RRPO update different from DPO? The gradient for $\mathcal{L}_{\text{RRPO}}$ can be obtained as:

$$\nabla_\theta \mathcal{L}_{\text{RRPO}} = -\nabla_\theta \mathbb{E}\Big[\log \sigma(u) + \alpha \cdot \mathbb{D}_{\text{TKL}}\big(x, y^+\big) \Big] = -\mathbb{E}\Big[\nabla_\theta \log \sigma(u) + \alpha \cdot \nabla_\theta \mathbb{D}_{\text{TKL}}\big(x, y^+\big) \Big]. \tag{7}$$

First, let's calculate the gradient for the ranking loss. Using chain rule, we can write $\nabla_\theta \log \sigma(u) = \Big[\frac{\sigma'(u)}{\sigma(u)} \nabla_\theta u \Big]$. Using the identity $\sigma'(u) = \sigma(u)(1 - \sigma(u))$, we have $\frac{\sigma'(u)}{\sigma(u)} = 1 - \sigma(u) = \sigma(-u)$. Therefore, $\nabla_\theta \log \sigma(u) = \Big[\sigma(-u) \nabla_\theta u \Big]$. Recalling our sub-sequence-level reward modeling defined in Equation 3, we can obtain:

$$\nabla_\theta u = \sum_{i=1}^{N} \nabla_\theta u_i = \beta \sum_{i=1}^{N} \Big(\sum_{j=s_i^+}^{e_i^+} \nabla_\theta \log \pi_\theta(t_j^+ | x, t_{<j}^+) - \sum_{j=s_i^-}^{e_i^-} \nabla_\theta \log \pi_\theta(t_j^- | x, t_{<j}^-) \Big), \tag{8}$$

as π_{ref} is fixed. Assuming the norm of the gradient of the log-probability for any single token is bounded, $\|\nabla_\theta \log \pi_\theta(t_j | x, t_{<j})\| \leq M$ for some $M > 0$, and that each differentiating sub-sequence y_i^+ or y_i^- has an average length L, we can further bound the gradient norm of the ranking loss:

$$\|\nabla_\theta \mathcal{L}_{\text{RRPO}}^{(\text{rank})}\| \leq \mathbb{E}\Big[\sigma(-u) \|\nabla_\theta u\| \Big] \leq \beta \sum_{i=1}^{N} (M L_i^+ + M L_i^-) \approx \beta M (2NL), \tag{9}$$

where L is the average length of the sub-sequences ($L \approx L_i^+ \approx L_i^-$). In contrast, the DPO gradient involves sums over the *entire* lengths of y^+ and y^-. A similar analysis for DPO yields a bound proportional to the total lengths:

$$\|\nabla_\theta \mathcal{L}_{\text{DPO}}\| \leq \beta M (|y^+| + |y^-|). \tag{10}$$

Crucially, the total length of the differentiating sub-sequences is typically much smaller than the total length of the full responses, i.e., $2NL \ll |y^+| + |y^-|$. Therefore, the upper bound on the gradient magnitude stemming from the ranking objective is smaller for RRPO compared to DPO:

$$\|\nabla_\theta \mathcal{L}_{\text{RRPO}}^{(\text{rank})}\| \ll \|\nabla_\theta \mathcal{L}_{\text{DPO}}\|.$$

Now, let's consider the term $\nabla_\theta \mathbb{D}_{\text{TKL}}(\cdot)$ in Equation 7. Based on the formulation of $\mathbb{D}_{\text{KL}}(\pi_{\text{ref}}||\pi_\theta) = \sum_a \pi_{\text{ref}}(a) \log \frac{\pi_{\text{ref}}(a)}{\pi_\theta(a)}$ where a represents a token in the vocabulary, and since π_{ref} fixed, we derive:

$$\nabla_\theta \mathbb{D}_{\text{TKL}}(x, y^+) = \sum_{t=1}^{|y^+|} \nabla_\theta \mathbb{D}_{\text{KL}}(\cdot) = -\sum_{t=1}^{|y^+|} \sum_a \pi_{\text{ref}}(a \mid x, y^+_{<t}) \, \nabla_\theta \log \pi_\theta(a \mid x, y^+_{<t}). \tag{11}$$

Note that $\nabla_\theta \mathbb{D}_{\text{TKL}}$ is always negative. Therefore, for $\alpha > 0$ in Equation 7, the gradient magnitude of $\mathcal{L}_{\text{RRPO}}$ is further reduced than $\mathcal{L}^{(\text{rank})}_{\text{RRPO}}$.

$$\boxed{\|\nabla_\theta \mathcal{L}_{\text{RRPO}}\| < \|\nabla_\theta \mathcal{L}^{(\text{rank})}_{\text{RRPO}}\| < \|\nabla_\theta \mathcal{L}_{\text{DPO}}\|}$$

This mathematical derivation confirms that the proposed RRPO loss function effectively reduces the gradient, as initially hypothesized. The reduced gradient of RRPO ensures more stable updates compared to DPO in gradient-based optimization while simultaneously enabling precise penalties on specific tokens without the risk of significant divergence. Furthermore, the $\mathbb{D}_{\text{TKL}}$ term acts as a trust-region constraint [35], preventing the model from making large, destabilizing updates. As a result, RRPO allows larger learning rates and yielding smoother convergence in practice. We present the pseudocode of RRPO in Appendix A.

4 Experiment Setup

Base models. We use VideoChat2$_{7B}$ [3], LLaVA-Video$_{7B}$ (also known as LLaVA-Next-Video$_{7B}$) [9], and LongVU$_{7B}$ [1] as our base models. These models are carefully selected to evaluate our method across diverse LLM architectures, vision encoders, cross-modal adapters, and training setups. For instance, among these models, VideoChat2 employs a video encoder, while the others rely on image encoders. Additionally, LongVU incorporates two vision encoders, whereas the rest use a single vision encoder. VideoChat2 utilizes QFormer [36] as its cross-modal adapter, whereas LLaVA-Video employs MLP projection layers. These models further differ in their LLM architectures, context lengths, and training setups, among other aspects.

Training data. Based on the availability and diversity of video-language instructions, we use VideoChat-IT [3] as our primary source for training samples. Specifically, we select a subset of VideoChat-IT encompassing eight video datasets: Kinetics700 [37], Something-Something-v2 [38], VideoChat [39], VideoChatGPT [40], CLEVRER [41], NEXTQA [42], EgoQA [43], and TGIF [44]. These datasets span a range of tasks, including video description, question answering, reasoning, and conversation. For the perturbation step, we mask a significant portion (25%-50%) of each frame and shuffle the temporal order. We explore three types of temporal perturbations: (*i*) random shuffling, (*ii*) local shuffling, and (*iii*) global shuffling. In random shuffling, frames are shuffled arbitrarily across time. For local shuffling, frames are initially segmented into smaller chunks, and the frames within each chunk are then shuffled. In global shuffling, the order of these chunks is shuffled, rather than individual frames. During inference, based on the LVLMs' input capacity, we utilize a maximum of 16, 64, and 100 frames for VideoChat2, LLaVA-Video, and LongVU, respectively.

Following the generation of the responses, we verify their correctness. For Multiple Choice Questions (MCQ) and Binary Question Answering (BinaryQA) tasks, verification is straightforward, using regex-based checks. However, for open-ended questions, this method proves inadequate, as semantically equivalent responses can be expressed in diverse phrasings. Consequently, for open-ended questions, we employ a powerful LLM, GPT-4o-mini [45], as a judge [46, 47, 48, 49], to ascertain correctness by comparing the generated response with the ground truth from the video instruction tuning dataset. Additionally, for long responses, we employ GPT-4o-mini to rewrite the correct response while incorporating the incorrect concepts from the generated response. This ensures that correct and incorrect concepts are aligned across both preferred and non-preferred responses. The key statistics of our training data are presented in Table 1. The prompts used during pre-processing with GPT-4o-mini are in Appendix D.

Table 1: Key statistics of the generated training samples.

Model	**# Samples**	**Unique pairs**
VideoChat2$_{7B}$	25K	18K
LongVU$_{7B}$	22K	14K
LLaVA-Video$_{7B}$	21K	16K

Implementation details. For all base models, we utilize LoRA [50] for training, applying it specifically to the LLMs while freezing all other parameters. Unless otherwise specified, we utilize

16 frames for self-alignment training; the rest follows the default training setup of each respective base model. We use $4\times$ A100 80GB GPUs for training, with the training time varying between 1 to 10 hours. Additional implementation details and hyperparameters are provided in Appendix E.

Evaluation benchmarks. To assess the impact of our self-alignment framework, we conduct evaluations across a diverse range of video understanding tasks. Specifically, we choose TVBench [17] and TempCompass [20] for fine-grained temporal understanding, VideoHallucer [19] and VidHalluc [51] for video hallucination, MVBench [3] and VideoMME [52] for short video understanding, and MLVU [24] and LongVideoBench [25] for long video understanding. It should be noted that these benchmarks, while selected for specific tasks, often assess overlapping video understanding capabilities. For instance, while VideoHallucer and VidHalluc are primarily used for hallucination detection, they also evaluate temporal grounding [19, 51]. Similarly, while TVBench mainly focuses on temporal understanding, it covers short video understanding as well [17]. Given its inclusion of short, medium, and long videos, VideoMME explores video understanding of varying lengths.

5 Results and Analysis

This section details our experimental evaluation, which encompasses a comprehensive and comparative analysis against other alignment methods and existing off-the-shelf aligned LVLMs. Furthermore, we present a detailed analysis focusing on the trade-offs between post-alignment divergence and performance. To gain deeper insights into the impact of our proposed method, we conduct experiments exploring its effects on fine-grained temporal understanding, hallucination mitigation, and performance on comprehensive video understanding of varying lengths. Finally, our evaluation includes an investigation into the influence of data, the scaling of input frames, and the presence of subtitles on the performance of our method. We conclude with a discussion regarding the generalization capabilities of our approach.

Table 2: Comparison with existing preference optimization methods and RRPO ablation variants. $\Delta = \frac{1}{N}\sum(\text{acc}_{\text{aligned}} - \text{acc}_{\text{base}})$ and $\%\Delta = \frac{100}{N}\sum(\text{acc}_{\text{aligned}} - \text{acc}_{\text{base}})/\text{acc}_{\text{base}}$, where N is the number of evaluation benchmarks used for each ablation. RRPO consistently outperforms existing alignment methods.

	TVB	**VHall**	**VMME**	**MLVU**	$\Delta/\%\Delta$
LongVU$_{\text{7B}}$ (base)	53.7	39.2	56.2	63.6	–
+ DPO [31]	54.3	40.9	56.6	63.6	0.7/1.5
+ DPA [27]	54.6	40.3	56.9	63.9	0.7/1.5
+ TDPO [34]	53.9	41.4	57.0	63.8	0.8/1.9
+ DDPO [53]	54.2	41.7	56.7	63.6	0.9/2.0
+ RRPO w/o R*	54.3	43.0	**57.8**	**64.5**	1.7/3.8
+ RRPO w/o $\mathbb{D}_{\text{TKL}}$	54.9	39.1	57.4	63.9	0.6/1.1
+ **RRPO (ours)**	**56.5**	**44.0**	57.7	**64.5**	**2.5/5.4**

(Abbreviations, TVB: TVBench; VHall: VideoHallucer; VMME: VideoMME)

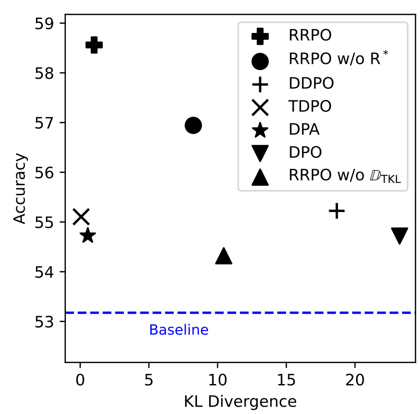


Figure 5: Performance relative to model divergence. RRPO exhibits the best performance-divergence trade-off.

Q1. How well does RRPO perform compared to other alignment methods? We conduct an in-depth analysis of RRPO against other recent preference optimization methods designed to address DPO's limitations, namely TDPO [34], DDPO [53], and DPA [27] and present the results in Table 2. More specifically, DDPO was introduced to provide fine-grained rewards, TDPO to enhance DPO's regularization, and DPA to address both of these challenges. A detailed discussion of their objectives, highlighting similarities and differences, is provided in Appendix B. Furthermore, we include two ablation variants of RRPO: one without the refined reward (RRPO w/o R*) and the other without the token-wise KL regularizer (RRPO w/o $\mathbb{D}_{\text{TKL}}$). Our study covers various aspects of video understanding, including fine-grained temporal understanding on TVBench, hallucination mitigation on VideoHallucer, comprehensive video understanding on VideoMME, and long video understanding on MLVU. The results presented in Table 2 demonstrate that RRPO consistently outperforms DPO, DPA, TDPO, and DDPO across all benchmarks. Moreover, among the ablation variants of RRPO, the inclusion of $\mathbb{D}_{\text{TKL}}$ alone yields significant performance gains, which are further enhanced by incorporating the refined reward. We present qualitative comparisons in Appendix G.

Table 3: Comparison with off-the-shelf aligned LVLMs. Ours LLaVA-Video-RRPO outperforms LLaVA-Video-TPO across all setups, both of which are based on LLaVA-Video$_{7B}$.

Model	**TV Bench**	**TempCompass$_{Avg}$**	**Video Hallucer**	**Vid Halluc**	**MV Bench**	**Video MME**	**MLVU$_{val}$ (M-Avg)**	**LongVideo Bench$_{Val}$**
LLaVA-Video-TPO [54]	51.1	66.1	50.6	76.3	60.6	**65.6**/71.5	68.7	60.1
LLaVA-Video-RRPO **(ours)**	**52.2**	**67.4**	**55.8**	**76.6**	**62.0**	65.5/**71.8**	**69.4**	**61.3**

Table 4: Evaluating our self-aligned LVLMs on diverse video understanding benchmarks. We **bold** the best in each group. 16f and 32f indicate the number of frames utilized during training. #F indicates number of frames used during inference. Here, we presents the overall results averaged across sub-categories where applicable, with detailed results available in Appendix C.

Models	**#F**	**TV Bench**	**TempCompass$_{Avg}$**	**Video Hallucer**	**Vid Halluc**	**MV Bench**	**Video MME**	**MLVU$_{val}$ (M-Avg)**	**LongVideo Bench$_{Val}$**
Video-LLaVA$_{7B}$ [10]		33.8	49.9	17.8	40.3	42.5	39.9	29.3	39.1
VideoLLaMA2$_{7B}$ [5]		42.9	43.4	10.0	66.3	54.6	62.4	48.4	–
LongVA$_{7B}$ [21]		–	57.0	–	–	–	52.6	42.1	–
MiniCPM-V 2.6$_{7B}$ [55]		–	66.3	–	–	–	60.9	–	54.9
NVILA$_{7B}$ [56]		–	–	–	–	68.1	64.2	70.1	57.7
LongVILA$_{7B}$ [8]		–	–	–	–	67.1	60.1	–	57.1
Qwen2-VL$_{7B}$ [57]		43.8	67.9	–	–	67.0	63.3	65.5	55.6
LLaVA-NeXT-Video-DPO$_{7B}$ [58]		–	53.8	32.0	–	–	–	–	43.5
ShareGPT4Video$_{8B}$ [59]		–	61.5	15.8	30.9	–	39.9	34.2	39.7
PLLaVA$_{13B}$ [60]		–	–	41.2	48.2	–	–	–	45.6
Aria$_{8x3.5B}$ [61]		51.0	69.6	–	–	69.7	67.6	–	63.0
LLaVA-NeXT-Video-DPO$_{34B}$ [58]		–	–	32.3	49.5	–	–	–	50.5
Qwen2-VL$_{72B}$ [57]		52.7	–	–	–	73.6	71.2	–	–
GPT-4o [62]		39.9	73.8	53.3	81.2	49.1	71.9	54.5	66.7
Gemini 1.5 Pro [11]		47.6	67.1	37.8	72.8	60.5	75.0	–	64.0
(The above models are presented for reference only, and may not be suitable for direct comparisons.)									
VideoChat2$_{7B}$ [3]		–	50.8	7.8	–	60.4	39.5	–	39.3
VideoChat2$_{7B}$	16	44.0	59.3	23.1	73.3	**60.2**	41.0	46.4	40.4
+ DPO (16f)	16	45.7	60.0	22.1	72.4	59.6	43.0	47.4	41.0
+ RRPO (16f) (Ours)	16	**45.8**	**60.2**	**32.9**	**76.4**	59.0	**44.3**	**47.9**	**42.8**
LLaVA-Video$_{7B}$ [9]		45.6	–	–	–	58.6	63.3	70.8	58.2
LLaVA-Video$_{7B}$	64	51.0	66.0	50.0	**76.6**	61.1	64.0	68.6	60.1
+ DPO (16f)	64	51.9	66.4	53.3	76.5	60.6	63.1	67.4	59.4
+ RRPO (16f) (Ours)	64	51.9	66.8	55.7	76.5	**62.2**	**64.5**	69.1	**60.4**
+ RRPO (32f) (Ours)	64	**52.2**	**67.4**	**55.8**	**76.6**	62.1	**64.5**	**69.4**	60.1
LongVU$_{7B}$ [1]		–	–	–	–	66.9	–	65.4	–
LongVU$_{7B}$	1fps	53.7	63.9	39.2	67.3	65.5	56.2	63.6	48.6
+ DPO (16f)	1fps	54.3	64.3	40.9	68.5	65.9	56.6	63.6	49.4
+ RRPO (16f) (Ours)	1fps	**56.5**	**64.5**	**44.0**	**71.7**	**66.8**	**57.7**	**64.5**	**49.7**

For transparency, we report base LVLMs' results from the literature where available, marked in grey. Differences from our reproduced results, both higher and lower, stem from variations in frame count, input prompt, GPT variant used for evaluation, and occasional implementation issues. For fairness, self-aligned variants are compared against our reproduced results.

Q2. How much does the model diverge post alignment? Understanding the extent of model divergence after alignment is essential to ensure that the process refines behavior without undermining core capabilities as excessive divergence can lead to instability, reduced generalization, and loss of valuable pre-trained knowledge. Figure 5 presents a comparative analysis of performance improvements against model divergence following preference optimization. Despite using a $10\times$ higher learning rate (which is facilitated through the smaller and more stable gradient updates), RRPO exhibits a KL divergence of 1 compared to DPO's KL divergence of 20. While TDPO and DPA exhibit almost the same divergence as RRPO, their performance across all evaluation benchmarks is substantially worse. RRPO, in contrast, demonstrates the optimal performance-divergence trade-off.

Q3. How does RRPO performance compared to off-the-shelf aligned LVLMs? We further evaluate our aligned LVLM against other off-the-shelf aligned LVLMs. Specifically, we compare our RRPO-trained LLaVA-Video with the concurrent work LLaVA-Video-TPO, both of which are based on the LLaVA-Video-Qwen$_{7B}$ [9]. LLaVA-Video-TPO is trained using a combination of DPO and SFT with a manually curated video-language dataset. The results in Table 3 demonstrate that RRPO is significantly more effective than concurrent methods. Our LLaVA-Video-RRPO outperforms LLaVA-Video-TPO across all setups, with performance gains of up to 5.2%.

Table 5: Impact of different spatio-temporal perturbations in generating non-preferred responses. Default perturbations for each model are colored.

	TVB	VHall	VMME	MLVU	$\Delta/\%\Delta$
VideoChat2$_{7B}$ (base)	44.0	23.1	41.0	46.4	–
+ `None`	40.7	16.0	39.9	43.8	-3.5/-11.6
+ `RS`	43.0	23.3	41.8	46.3	0.0/0.1
+ `Mask`	44.0	26.2	43.4	48.8	2.0/6.1
+ `LS-Mask`	44.6	28.2	44.6	49.4	3.1/9.7
+ `GS-Mask`	44.2	28.8	44.1	46.3	2.2/8.1
+ `RS-Mask`	45.8	32.9	44.3	47.9	**4.1/14.4**
LLaVA-Video$_{7B}$ (base)	51.0	50.0	64.0	68.6	–
+ `LS-Mask`	51.9	55.7	64.5	69.1	**1.9/3.7**
+ `GS-Mask`	51.3	54.8	64.2	68.7	1.4/2.7
+ `RS-Mask`	51.5	51.6	64.6	69.6	0.9/1.6
LongVU$_{7B}$ (base)	53.7	39.2	56.2	63.6	–
+ `LS-Mask`	56.5	44.0	57.7	64.5	**2.5/5.4**
+ `GS-Mask`	55.0	43.9	56.9	64.5	1.9/4.3
+ `RS-Mask`	55.1	42.4	57.0	64.3	1.5/3.3

Table 6: Impact of data size.

	TVB	VHall	VMME	MLVU	$\Delta/\%\Delta$
Baseline	51.0	50.0	64.0	68.6	–
+ 5K	50.9	53.7	64.0	69.0	1.0/1.9
+ 10K	51.2	53.8	64.3	69.0	1.2/2.3
+ 15K	51.8	54.4	64.2	68.9	1.4/2.8
+ 20K	51.9	55.7	64.5	69.1	**1.9/3.7**

Table 7: Impact of varying the number of frames at inference.

		TVB	MVB	LVB
32	Baseline	49.6	61.2	58.4
	+ RRPO	51.3	61.7	58.9
64	Baseline	51.0	61.1	60.1
	+ RRPO	**52.2**	**62.1**	60.1
128	Baseline	49.4	60.5	60.3
	+ RRPO	51.3	61.2	**61.3**

Table 8: Impact of using subtitles along with videos (VMME).

	without	with
VideoChat2$_{7B}$	41.0	48.0
+ RRPO	**44.3**	**49.4**
LLaVA-Video$_{7B}$	63.8	67.4
+ RRPO	**64.5**	**68.0**
LongVU$_{7B}$	56.2	62.0
+ RRPO	**57.7**	**63.1**

(Abbreviations, `RS`: Random Shuffle; `LS`: Local Shuffle; `GS`: Global Shuffle; TVB: TVBench; VHall: VideoHallucer; VMME: VideoMME; LVB: LongVideoBench.)

Q4. Does our method improve fine-grained temporal understanding? To evaluate this, we utilize TVBench [17] and TempCompass [20], designed to test the temporal understanding abilities of LVLMs. TVBench tests capabilities across various temporal tasks, including action localization, directional movement, and scene transitions, among others. Similarly, TempCompass evaluates performance on video captioning, caption matching, MCQ, and BinaryQA, covering video understanding tasks such as event ordering, action identification, and state change. As shown in Table 4, our method, RRPO, consistently improves base model performance by up to 2.8%, demonstrating its effectiveness in enhancing fine-grained temporal understanding and outperforming DPO across all setups.

Q5. Does our method effectively mitigate hallucinations? Hallucination occurs when LVLMs produce responses that are ungrounded, referred to as intrinsic hallucination, or unverifiable, referred to as extrinsic hallucination. Hallucination presents a significant obstacle to the reliable use of LVLMs. To evaluate the impact of our method on video hallucination, we employ VideoHallucer [19] and VidHalluc [51]. VideoHallucer tests LVLMs for both extrinsic and intrinsic hallucinations while including both spatial and temporal hallucinations. Additionally, VidHalluc focuses specifically on temporal hallucinations, such as action hallucination. As shown in Table 4, RRPO significantly reduces hallucination across all base models. Specifically, RRPO improves performance by 4.8% to 8.8% on VideoHallucer and by up to 4.4% on VidHalluc. In most cases, RRPO demonstrates a substantial performance advantage over DPO, with gains reaching 10.8%.

Q6. Does our method improve comprehensive video understanding across varying video lengths? To evaluate the comprehensive video understanding capabilities of LVLMs across varying video lengths, we leverage four benchmarks: MVBench [3], VideoMME [52], MLVU [24], and LongVideoBench [25]. Together, these benchmarks span a wide variety of perception and reasoning tasks, focusing on objects, actions, their attributes, and holistic video understanding. Among these, MVBench is specifically designed for a comprehensive evaluation of *short videos*, while MLVU and LongVideoBench offer thorough evaluations for *long videos*. VideoMME provides a comprehensive assessment across videos of *varying lengths*. As shown in Table 4, consistent improvements are observed across all benchmarks for LongVU and LLaVA-Next. For VideoChat2, self-alignment leads to substantial gains in three out of four setups, with only a minor regression in MVBench. Furthermore, RRPO consistently outperforms DPO, further demonstrating the advantages of fine-grained alignment in enhancing comprehensive video understanding.

Q7. How do perturbations in our data generation pipeline impact the quality of non-preferred responses? To assess the impact of the perturbations on the quality of non-preferred responses, we conduct in-depth analyses and present the results in Table 5. Our key observations are as follows: First, non-preferred responses generated without video perturbations leads to diminished model performance, likely due to reduced generalizability. Second, temporal perturbations alone are ineffective, although their combination with masking significantly boosts performance. Among the spatio-temporal perturbations, random shuffling with masking (`RS-Mask`) improves performance for models processing fewer frames (e.g., VideoChat2) whereas local shuffling with masking (`LS-Mask`) proves superior for models handling longer sequences (e.g., LongVU, LLaVA-Video).

Q8. How does our method scale with training data? To test the impact of scaling the amount of data, we perform an experiment by varying the number of training samples from 5K to 20K, incrementing by 5K. As shown in Table 6, performance improves with data size. This suggests that our data-generation pipeline is effective in producing high-quality training samples for self-alignment.

Q9. How does performance vary with the number of input frames? We investigate the effect of scaling the number of visual input frames during both inference and self-alignment training. For inference, we evaluate LLaVA-Video using 32, 64, and 128 frames across TVBench, MVBench, and LongVideoBench. The results are presented in Table 7. Our key observations are as follows: First, RRPO consistently improves performance over the base model across all setups. Second, neither the base model nor RRPO exhibits performance gains beyond 64 frames on TVBench and MVBench. This is likely due to the short-video nature of these benchmarks, where higher frame counts result in redundant frame resampling. However, for long-video understanding, increasing frame counts yields performance improvements, particularly for RRPO with a 1.2% improvement compared to the base model's 0.2% gain. Subsequently, we explore the impact of increasing the number of frames during self-alignment training. Specifically, we raise the frame count from our default 16 to 32. The results presented in Table 4 demonstrate that this increase enhances performance. Notably, we observe a consistent performance improvement on fine-grained temporal understanding tasks, as evidenced by the gains on TVBench and TempCompass.

Q10. Does our method retain its performance advantage with subtitles? Subtitles generally enhance video understanding by providing additional language cues that LVLMs can leverage. Thus, we investigate whether our method maintains its benefits over base models when subtitles are included. As shown in Table 8, our method demonstrates consistent improvements across all setups.

Q11. Does our method generalize across diverse LVLM architectures and training setups? Given the rapid evolution of LVLMs, we carefully select VideoChat2, LLaVA-Video, and LongVU as the base models to cover a wide variety of design choices and training methodologies (e.g., Table 4). Specifically, VideoChat2 uses the UMT [63] video encoder, while LLaVA-Video and LongVU use DINOv2 [64] and SigLIP [65] as their image encoders. On the other hand, LongVU employs dual vision encoders, unlike the others. Moreover, the cross-modal adapters range from query-based for VideoChat2 to MLP projections for LLaVA-Video, and a combinations of both in the case of LongVU. Furthermore, the overlap between self-alignment training samples and instruction tuning data differs across models, with VideoChat2 having the highest overlap and LLaVA-Video having the least. Importantly, our self-alignment consistently improves performance across these diverse setups, even when reusing instruction tuning data.

6 Related work

Recent years have seen a surge in the development of LVLMs with improved video understanding capabilities [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. This progress can be attributed to several key factors: (1) the development of diverse video benchmarks [3, 7, 66, 43, 67], which enable LVLMs to follow human instructions and reason across a variety of video tasks; (2) architectural innovations that support the use of rich, dense visual features and enable efficient processing of long sequences [21, 8, 1, 68, 69, 70, 5, 71]; and (3) advancements in training algorithms for both the pre-training [72, 63, 65, 73] and post-training [74, 75, 31] stages. LVLM training generally involves multi-stage processes [76, 77], with pre-training typically focusing on representational alignment between video and language [3, 78, 36], and post-training refining the model's ability to follow instructions [3, 78, 79], reduce hallucination [53, 27], improve reasoning skills [80, 81], and align the model with human preferences [80, 81]. Our introduced RRPO is a post-training alignment method designed to enhance the overall video understanding capabilities of LVLMs. While concurrent works [82, 83, 54] have explored post-training alignments of LVLMs, they directly adopt DPO [31], which proves ineffective in fine-grained alignment, as discussed in Section 5.

7 Conclusion

To improve the video understanding abilities of LVLMs, we design a self-alignment framework that enables LVLMs to learn from their own errors. These errors commonly occur due to their lack of spatio-temporal reasoning, over-reliance on linguistic cues, and spurious correlations between co-occurring concepts, among others. To effectively align LVLMs against such errors, we introduce RRPO, a novel preference optimization method designed for fine-grained alignment through refined reward modeling with strong regularization. Our in-depth experiments and analyses show that RRPO training is more stable and highly effective compared to prior and concurrent preference optimization methods across diverse tasks. Moreover, the fine-grained reward modeling in RRPO improves capabilities without causing significant divergence from the base models. Our proposed self-alignment with RRPO exhibits consistent improvements across all setups over the base models, effectively reducing hallucination and improving spatio-temporal reasoning, thus enabling safer and more reliable use of LVLMs. Moreover, we show that the approach scales well with more data and high-resolution temporal inputs, and generalizes well across diverse LVLM architectures and training setups. Future work can further investigate iterative self-alignment methodologies with RRPO, moving beyond the static dataset used in this work.

Acknowledgment

We thank Ahmad Beirami for his valuable feedback and suggestions, which helped improve our paper. We also thank the Bank of Montreal and Mitacs for funding this research, and the Vector Institute for providing computational resources.

References

[1] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 1, 5, 7, 9

[2] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024. 1, 9

[3] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 1, 5, 6, 7, 8, 9

[4] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 1, 9

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 7, 9

[6] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1, 9

[7] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024. 1, 9

[8] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1, 7, 9

[9] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 5, 7, 9

[10] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 7, 9

[11] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jeanbaptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 7, 9

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 1, 9

[13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 9

[14] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2, 2023. 1, 9

[15] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 9

[16] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024. 1, 2

[17] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 1, 2, 6, 8

[18] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *ACCV*, pages 18–34, 2024. 1, 2

[19] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024. 1, 6, 8

[20] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 1, 6, 8

[21] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1, 2, 7, 9

[22] Junqi Ge, Ziyi Chen, Jintao Lin, Jinguo Zhu, Xihui Liu, Jifeng Dai, and Xizhou Zhu. V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding. *arXiv preprint arXiv:2412.09616*, 2024. 1, 2

[23] Hongchen Wei and Zhenzhong Chen. Visual context window extension: A new perspective for long video understanding. *arXiv preprint arXiv:2409.20018*, 2024. 1, 2

[24] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 1, 6, 8

[25] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *NeurIPS*, 37:28828–28857, 2025. 1, 6, 8

[26] Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*, 2025. 2

[27] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Data-augmented phrase-level alignment for mitigating object hallucination. *arXiv preprint arXiv:2405.18654*, 2024. 2, 3, 4, 6, 9, 17

[28] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 2

[29] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 2

[30] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *NeurIPS*, 36:2511–2565, 2023. 2

[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36:53728–53741, 2023. 2, 3, 6, 9, 16

[32] Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, Jian Xie, Zhijie Deng, and Dong Yan. 3d-properties: Identifying challenges in dpo and charting a path forward. *arXiv preprint arXiv:2406.07327*, 2024. 3

[33] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024. 3

[34] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024. 4, 6, 17

[35] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, pages 1889–1897. PMLR, 2015. 5

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5, 9

[37] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 5

[38] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017. 5

[39] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 5

[40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 5

[41] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 5

[42] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 5

[43] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 5, 9

[44] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 5

[45] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[46] Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. 5

[47] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 5

[48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36:46595–46623, 2023. 5

[49] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023. 5

[50] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5, 21

[51] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. *arXiv preprint arXiv:2412.03735*, 2024. 6, 8

[52] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 6, 8

[53] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, pages 13807–13816, 2024. 6, 9, 16

[54] Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025. 7, 9

[55] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 7

[56] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024. 7

[57] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7

[58] Y Zhang, B Li, H Liu, Y Lee, L Gui, D Fu, J Feng, Z Liu, and C Li. Llava-next: A strong zero-shot video understanding model. 2024. 7

[59] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *NeurIPS*, 37:19472–19495, 2024. 7

[60] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 7

[61] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 7

[62] 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. 7

[63] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *ICCV*, pages 19948–19960, 2023. 9

[64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 9

[65] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 9

[66] Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision. *arXiv preprint arXiv:2407.06189*, 2024. 9

[67] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 9

[68] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, pages 18221–18232, 2024. 9

[69] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 9

[70] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *ECCV*, pages 202–218. Springer, 2024. 9

[71] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 9

[72] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 9

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021. 9

[74] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017. 9

[75] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 9

[76] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024. 9

[77] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 9

[78] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 9

[79] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 9

[80] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 9

[81] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025. 9

[82] Yogesh Kulkarni and Pooyan Fazli. Videosavi: Self-aligned video language models without human supervision. *arXiv preprint arXiv:2412.00624*, 2024. 9

[83] Daechul Ahn, Yura Choi, San Kim, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. i-srt: Aligning large multimodal models for videos by iterative self-retrospective judgment. *arXiv preprint arXiv:2406.11280*, 2024. 9

[84] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training. In *USENIX ATC*, pages 551–564, 2021. 21

[85] Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *SC*, pages 1–14, 2021. 21

Appendix

A RRPO Pseudocode (PyTorch Style)

```
import torch
import torch.nn.functional as F

def rrpo_loss(self, logits, ref_logits, phrase_ids, alpha, beta):
    """
    logits:         logits from pi_theta.
                    shape: (batch_size, sequence_length, vocab_size)
    ref_logits:     logits from pi_ref.
                    shape: (batch_size, sequence_length, vocab_size)
    phrase_ids:     phrase identifiers where tokens belonging to the same
                    phrase share the same value; additionally, they remain
                    the same between correct and misaligned phrases to
                    maintain correspondence.
                    shape: (batch_size, sequence_length)
    correct_idx:    indices of the correct responses in batch
    wrong_idx:      indices of the wrong responses in batch
    alpha:          coefficient to control the token-wise KL divergence
    beta:           coefficient to control reward/penalty
    """

    # compute log probabilities
    logps = logits.log_softmax(dim=-1)
    ref_ps = ref_logits.softmax(dim=-1)
    ref_logps = ref_ps.log()

    # compute token-wise KL divergence
    token_wise_kl = (ref_ps * (ref_logps - logps)).sum(dim=-1)

    # compute the margin
    logps_margin = logps - ref_logps

    # accumulate log probabilities of the phrases with key concepts
    # here 0 indicates ignored tokens
    unique_phrase_ids = torch.unique(phrase_ids, sorted=True)[1:]
    phrase_logps_margin = torch.zeros(
            phrase_ids.size(0), len(unique_phrase_ids))
    for i, phrase_id in enumerate(unique_phrase_ids):
        mask = (phrase_ids == phrase_id).float()
        phrase_logps_margin[:, i] = (logps_margin * mask).sum(dim=-1)

    chosen_logps_margin = phrase_logps_margin[correct_idx]
    rejected_logps_margin = phrase_logps_margin[wrong_idx]
    logits_margin = (chosen_logps_margin -
                        rejected_logps_margin).sum(dim=-1)

    chosen_token_wise_kl = token_wise_kl.sum(dim=-1)[chosen_idx]

    losses = -torch.logsigmoid(beta * logits_margin)
             + alpha * chosen_token_wise_kl

    return losses
```

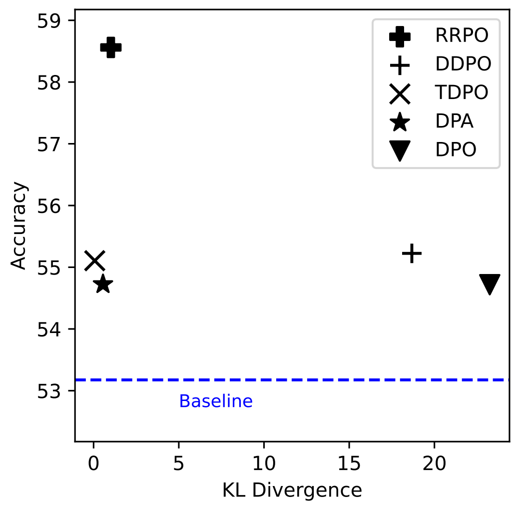


Figure S1: Comparison of characteristics among preference optimization methods. Both DDPO and DPO diverge significantly from their initial state after alignment. While TDPO and DPA are effective in restricting model divergence, they are less effective in improving performance. RRPO achieves excellent performance with minimal divergence from the base model.

B Comparing Loss Formulation of RRPO and Other Methods

This section presents a comparative analysis between RRPO and other preference optimization methods studied in this work, i.e., DPO, DDPO, TDPO, and DPA. We begin by outlining the respective loss functions, followed by a detailed discussion of their similarities and differences.

Comparison with DPO.

As previously discussed, the DPO [31] loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) &= -\mathbb{E}\Big[\log\sigma\big(r_\theta(x,y^+) - r_\theta(x,y^-)\big)\Big],\\ &= -\mathbb{E}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta\log\frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\Big)\Big].\end{aligned} \tag{S1}$$

The DPO loss function calculates reward over all tokens in y^+ and y^-, despite the fact that there might be a few sub-sequences that are conceptually different. This approach results in coarse-grained reward modeling, and because it penalizes all tokens in the response, the loss function accumulates a large gradient and tends to diverge significantly from the base model, potentially resulting in weak alignment, as shown in Figure S1. RRPO is introduced to address two key challenges of DPO: to provide fine-grained feedback and restrict the divergence of the model away from its initial state.

Comparison with DDPO.

DDPO [53] extends DPO by incorporating a weighted reward based on the sub-sequence level differences between y^+ and y^-, and is defined as:

$$\mathcal{L}_{\text{DDPO}}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta\log\frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\Big)\Big], \tag{S2}$$

where $\log\pi(y|x) = \sum_{y_i\in y}\log p(y_i|x,y_{<i})$ is modified as:

$$\log\pi(y|x) = \frac{1}{N}\big[\sum_{y_i\in y_{\text{same}}}\log p(y_i|x,y_{<i}) + \gamma\sum_{y_i\in y_{\text{different}}}\log p(y_i|x,y_{<i})\big].$$

Here, y_{same} and $y_{\text{different}}$ indicate unchanged and changed segments between y^+ and y^-. Moreover, $\gamma > 1$ is a weighting hyperparameter, and larger γ indicates more weight on those changed segments.

While DDPO is designed to provide fine-grained feedback, its loss formulation is not as effective as RRPO and is also prone to diverging far from its initial state, similar to DPO, due to weak regularization, see Figure S1.

Comparison with TDPO.

TDPO [34] is also derived from DPO, incorporating an additional regularization term ($\mathbb{D}_{\text{TKL}}$) between π_θ and π_{ref}, which is defined as:

$$\begin{aligned}\mathcal{L}_{\text{TDPO}}(\pi_\theta;\pi_{\text{ref}}) = -\,\mathbb{E}\Bigg[\log\sigma\Bigg(&\big(r_\theta(x,y^+) - r_\theta(x,y^-)\big)\\ &-\alpha\Big(\beta\mathbb{D}_{\text{TKL}}\left(x,y_w;\pi_{\text{ref}}\|\pi_\theta\right) - sg\big(\beta\mathbb{D}_{\text{TKL}}\left(x,y_c;\pi_{\text{ref}}\|\pi_\theta\right)\big)\Big)\Bigg)\Bigg].\end{aligned} \tag{S3}$$

As shown in Figure S1, TDPO is effective in restricting the divergence of the base model, but its performance is almost the same as DPO and falls short of our RRPO.

Comparison with DPA.

DPA [27] is a phrase-level alignment method unlike DPO and its variants. The DPA loss is composed of two terms where the first term computes the relative log-probability between two phrases of y^+ and y^-, and the second term works as a regularizer between π_θ and π_{ref}, formulated as:

$$\mathcal{L}_{\text{DPA}}(\pi_\theta;\pi_{\text{ref}}) = -\,\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} -\log\frac{P(y_i^+)}{P(y_i^+)+P(y_i^-)} + \alpha\cdot\mathbb{D}_{\text{TKL}}\big(x,y^+;\pi_{\text{ref}}\|\pi_\theta\big)\right], \tag{S4}$$

where $P(y_i^+)$ and $P(y_i^-)$ denote the probability of i-th phrase in y^+ and y^-. Assume, y is expressed as a sequence of tokens $\{t_1, t_2, \ldots, t_{|T|}\}$, then the probability of i-th phrase can be computed as $\prod_{j=s_i}^{e_i} \pi_\theta(t_j|x, t_{<j})$, where s_i and e_i represent the start and end token indices.

The loss formulation of RRPO draws inspiration from DPA to achieve fine-grained alignment without the risk of divergence. However, we identify a fundamental limitation in DPA: it directly adjusts the probabilities of π_θ to modify the probability ratio between preferred and non-preferred phrases. This approach leads to an inaccurate probability ratio after the initial pair of preferred and non-preferred segments, as subsequent segment probabilities become dependent on their preceding elements. Therefore, DPA is not accurate in providing fine-grained feedback for sequences composed of multiple sub-sequences of key concepts. As shown in Figure S1, DPA, while successful in controlling divergence, is ineffective in improving performance.

C Additional Results

Statistical analysis between RRPO and DPO. We perform statistical analysis between RRPO and DPO aligned model. We consider the performance (denoted as `Score`) difference, whether an improvement or decline, of the `Model1` relative to the `Model2` on a specific task to be statistically significant if the Adjusted Δ exceeds zero, where Δ denotes the difference in performance between the `Model1` and `Model2`. Statistical significance is assessed using the Standard Error (SE) at a 95% confidence level. The corresponding mathematical formulations are presented below.

$$\mathrm{SE} = \sqrt{\frac{\texttt{Score}_{\texttt{Model1}} \times (1 - \texttt{Score}_{\texttt{Model2}})}{\text{number of samples}}},$$

$$\Delta = \texttt{Score}_{\texttt{Model1}} - \texttt{Score}_{\texttt{Model2}},$$

$$\text{Adjusted } \Delta = \Delta - 1.96 * \mathrm{SE}.$$

The results presented in Table S1 shows that RRPO almost always outperforms DPO and even in several cases improvements are statistically significant.

Table S1: Statistical analysis between RRPO and DPO variants. Green underline indicates that the performance improvements are statistically significant. Also, there are no instances in which the DPO variants achieve statistically significant gains over RRPO.

Model	**TV Bench**	**TempCompass$_{\text{Avg}}$**	**Video Hallucer**	**Vid Halluc**	**MV Bench**	**Video MME**	**MLVU$_{\text{val}}$ (M-Avg)**	**LongVideo Bench$_{\text{Val}}$**
VideoChat2$_{7B}$+DPO	45.7	60.0	22.1	72.4	59.7	43.0	47.4	41.0
VideoChat2$_{7B}$+RRPO	45.8	60.2	<u>32.9</u>	<u>76.4</u>	59.1	44.3	47.9	42.8
LLaVA-Video$_{7B}$+DPO	51.9	66.4	53.3	76.5	60.2	63.1	67.4	59.4
LLaVA-Video$_{7B}$+RRPO	51.9	66.8	55.7	76.5	62.0	64.5	69.1	60.4
LLaVA-Video$_{7B}$+TPO	51.1	66.1	50.6	76.3	60.6	65.6	68.9	60.1
LLaVA-Video$_{7B}$+RRPO	52.2	67.4	<u>55.8</u>	76.6	62.0	65.5	69.4	61.3
LongVU$_{7B}$+DPO	54.3	64.3	40.9	68.5	65.8	56.6	63.6	49.4
LongVU$_{7B}$+RRPO	<u>56.5</u>	64.5	<u>44.0</u>	<u>71.7</u>	66.7	57.7	64.5	49.7

Detailed results. This section details the results for the subcategories of the evaluation benchmarks used in our study.

Table S2: Detailed results on TempCompass.

Models	Caption Matching	Captioning	Multi-choice	Yes-No	Avg.
VideoChat2$_{7B}$	69.5	46.6	58.0	63.0	59.3
+ RRPO	73.2	48.5	56.6	62.6	60.2
LlavaVideo$_{7B}$	75.1	50.2	67.6	71.0	66.0
+ RRPO	75.8	52.0	68.6	70.9	66.8
+ RRPO (32f)	76.6	53.0	68.7	71.3	67.4
LongVU$_{7B}$	74.7	49.8	63.9	67.3	63.9
+ RRPO	75.2	50.6	64.7	67.4	64.5

Table S3: Detailed results on VideoHallucer.

Model	Object relation	Temporal	Semantic	Factual	Non-factual	Avg.
VideoChat2$_{7B}$	47.5	8.0	38.5	1.0	20.5	23.1
+ RRPO	53.5	24.0	55.0	5.0	27.0	32.9
LlavaVideo$_{7B}$	66.0	56.5	65.5	13.5	48.5	50.0
+ RRPO	65.5	65.5	71.0	23.5	53.0	55.7
+ RRPO (32f)	65.5	65.5	71.5	23.5	53.0	55.8
LongVU$_{7B}$	50.5	46.0	43.0	17.0	39.5	39.2
+ RRPO	53.0	48.0	50.0	26.0	43.0	44.0

Table S4: Detailed results on VidHalluc.

Model	BinaryQA	MCQ	Scene Transition	Avg.
VideoChat2$_{7B}$	66.8	84.9	68.2	73.3
+ RRPO	72.7	85.5	70.9	76.4
LlavaVideo$_{7B}$	77.9	91.4	60.6	76.6
+ RRPO	78.4	91.6	59.5	76.5
+ RRPO (32f)	78.6	91.7	59.5	76.6
LongVU$_{7B}$	71.4	87.0	43.4	67.3
+ RRPO	74.2	88.2	52.7	71.7

Table S5: Detailed results on VideoMME.

Model	Short	Medium	Long	Avg.
VideoChat2$_{7B}$	49.0	38.6	35.6	41.0
+ RRPO	52.2	41.9	38.8	44.3
LlavaVideo$_{7B}$	76.3	62.8	52.8	64.0
+ RRPO	76.6	63.1	53.8	64.5
+ RRPO (32f)	76.7	62.9	53.9	64.5
LongVU$_{7B}$	66.1	54.7	47.9	56.2
+ RRPO	67.7	55.0	50.3	57.7

D Additional Details of Training Data

Prompt templates.

The instructions used in processing open-ended generated responses employing GPT4o are presented in Figures S2 and S3.

```
% {Prompt used in open-ended response processing stage 1}
Thoroughly read the question and the given answers.

Your task is to determine whether the "Predicted answer" is "Correct" or "Wrong"
based on the "Question" and "Reference answer".
To determine correctness, focus on the key aspects in the answers, such as
objects, actions, and their attributes, among others.
The "Predicted answer" may have partial information in comparison to the
"Reference answer", in that case check whether at least the partial information
can be fully verified based on the "Reference answer".

Please respond with any of the following and nothing else:

- "Correct" if the predicted answer is correct based on the reference answer.
- "Wrong" if the predicted answer is not fully correct based on the reference
answer.
- "Undecided" if you are not sure about their correctness.

Question: {question}
Reference answer: {ground_truth}
Predicted answer: {generated_response}
```

Figure S2: Prompt used in open-ended response processing stage 1.

```
% {Prompt used in open-ended response processing stage 2}
***Turn 1***

Identify the key differences between these two sentences.
To identify differences focus on the key aspects in the sentences,
such as objects, actions, and their attributes, among others.
If there are no key difference between these two sentences,
please respond with "None" and nothing else.

Sentence 1: {sentence_from_ground_truth}
Sentence 2: {sentence_from_generated_response}

***Turn 2***

Please rewrite the "Sentence 1" by incorporating the differences you
mentioned earlier. Your final response should contain only the revised
sentence and nothing else.

Sentence 1: {sentence_from_ground_truth}
```

Figure S3: Prompt used in open-ended response processing stage 2.

E Implementation Details

Training hyperparameters.

Table S6: Details of training hyperparameters.

	VideoChat2	LLaVA-Video	LongVU
LLM	Mistral	Qwen2	Qwen2
Vision encoder	UMT	SigLIP	SigLIP+DINOv2
Trainable module	LoRA in LLM and everything else is kept frozen		
LoRA setup [50]	rank=128, alpha=256		
Learning rate	2e-5	5e-6	5e-6
Learning rate scheduler	Cosine	Cosine	Cosine
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.02	0.0	0.0
Warmup ratio	-	0.03	0.03
Epoch	1	1	1
Batch size per GPU	2	1	1
Batch size (total)	32	32	32
α (loss coefficient)	0.01	0.01	0.05
β (loss coefficient)	0.9	0.1	0.5
Memory optimization	-	Zero stage 3 [84, 85]	FSDP

Licenses of existing assets used.

- VideoChat2 (Apache License 2.0): *https://huggingface.co/OpenGVLab/VideoChat2_stage3_Mistral_7B*
- LLaVA-Video (Apache License 2.0): *https://huggingface.co/lmms-lab/LLaVA-Video-7B-Qwen2*
- LongVU (Apache License 2.0): *https://huggingface.co/Vision-CAIR/LongVU_Qwen2_7B*
- VideoChat-IT (MIT): *https://huggingface.co/datasets/OpenGVLab/VideoChat2-IT*

F Broader impact

As generative models are increasingly deployed in real-world applications, there is a growing need for post-training methods that enable fine-grained alignment with human preferences and values. Our proposed method, RRPO, can be applied to align generative models in both language and multimodal settings. By facilitating more precise alignment, RRPO has the potential to improve the safety, reliability, and usability of these models for real-world usage.

G Qualitative Results

We present several examples in Figures S4 - S7 highlighting the effectiveness of RRPO over the base model and other preference optimization methods (e.g., DPO) in diverse video understanding tasks. We also present some failure cases in Figures S8 - S9.

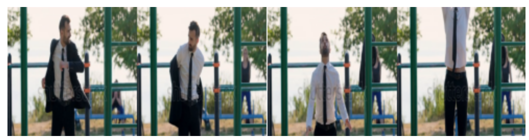
Question: The person uses multiple similar objects to play an occlusion game. Where is the hidden object at the end of the game from the person's point of view?
Options:
(A) Under the third object from the left.
(B) Under the second object from the left.
(C) Under the first object from the left.
VideoChat2: (A) Under the third object from the left.
VideoChat2 + DPO: (B) Under the second object from the left.
VideoChat2 + RRPO: (C) Under the first object from the left.

Question: Choose the caption that closely matches the video.
VideoChat2: The video shows a cat licking its paw and then two people lifting weights.
VideoChat2 + DPO: The video shows a cat licking its paw and then two people lifting weights.
VideoChat2 + RRPO: The video shows a cat licking its paw after two people lift weights.

Question: What did the person do first?
Options:
(A) Held the phone/camera.
(B) Threw the clothes.
LLaVA-Video: (A) Held the phone/camera.
LLaVA-Video + DPO: (A) Held the phone/camera.
LLaVA-Video + RRPO: (B) Threw the clothes.

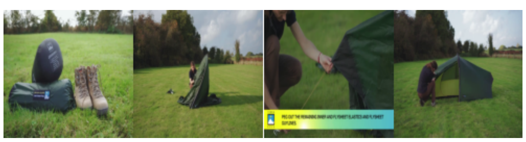
Question: Choose the caption that closely matches the video.
LLaVA-Video: The man above is jumping rope at a steady pace.
LLaVA-Video + DPO: The man above is jumping rope at a steady pace.
LLaVA-Video + RRPO: The man above is jumping rope at the same speed.

Question: What's the right option for how the scenes in the video change?
Options: (A) From the elevator to the secretary's office.
(B) From the secretary's office to the elevator.
LongVU: (A) From the elevator to the secretary's office.
LongVU + DPO: (A) From the elevator to the secretary's office.
LongVU + RRPO: (B) From the secretary's office to the elevator.



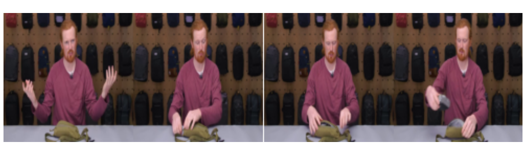
Question: Choose the caption that closely matches the video.
LongVU: A man putting on his jacket and then jumping up to grab the bar.
LongVU + DPO: A man putting on his jacket and then jumping up to grab the bar.
LongVU + RRPO: A man taking off his jacket and then jumping up to grab the bar.

Figure S4: Qualitative examples on fine-grained temporal understanding tasks.

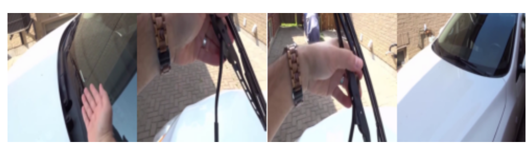


Question: Based on the video, should we set up the sandwiches in a tent when we pitch a tent?
VideoChat2: Yes, it is recommended to set up the sandwiches in a tent when pitching a tent.
VideoChat2 + DPO: Yes, it is recommended to set up the sandwiches in a tent when pitching a tent.
VideoChat2 + RRPO: The video does not provide any information about setting up sandwiches in a tent.

Question: Based on the video, should we rub back and forth with sandpaper or rust remover when we clean rusty pot?
VideoChat2: Yes, it is recommended to rub back and forth with sandpaper or rust remover when cleaning a rusty pot.
VideoChat2 + DPO: To clean a rusty pot, it is recommended to use sandpaper and rub.
VideoChat2 + RRPO: No, according to the video, we should not use sandpaper or rust remover to clean a rusty pot.



Question: Does the person in the video unzip the bag?
LLaVA-Video: Yes
LLaVA-Video + DPO: Yes
LLaVA-Video + RRPO: No

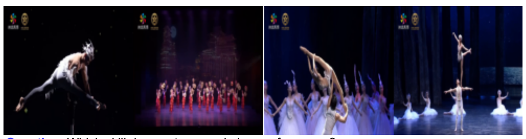


Question: Based on the video, should we wipe the front windshield when we replace a wiper head?
LLaVA-Video: Yes
LLaVA-Video + DPO: Yes
LLaVA-Video + RRPO: No

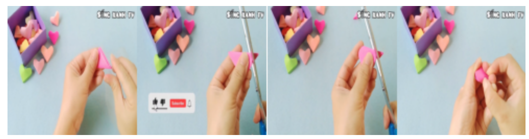
Question: Is there a women sitting at a chair in the video?
LongVU: Yes, there is a woman sitting on a chair.
LongVU + DPO: Yes, there is a woman sitting on a chair.
LongVU + RRPO: No, there is no woman sitting on a chair.

Question: Based on the video, should we snip the wire when we make RJ45 cable?
LongVU: Yes, snipping the wires evenly ensures a proper fit.
LongVU + DPO: Yes, snipping the wires evenly ensures a proper fit.
LongVU + RRPO: No, we should not snip the wire.

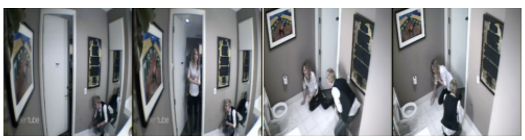
Figure S5: Qualitative examples on video hallucination tasks.



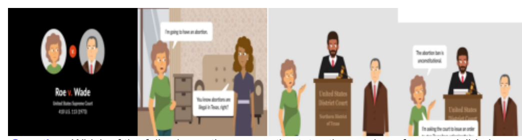
Question: Which skill does not appear in her performance?
(A) Pole skills.
(B) Loop skills.
(C) Stilts skills.
(D) Human pyramid skills.
VideoChat2: (A) Pole skills.
VideoChat2 + DPO: (D) Human pyramid skills.
VideoChat2 + RRPO: (C) Stilts skills.



Question: What is the function of the scissors that are used in the video?
(A) Making the paper look more like a heart.
(B) Destroying the paper.
(C) Cutting the paper into pieces.
(D) None of the above.
VideoChat2: (C) Cutting the paper into pieces.
VideoChat2 + DPO: (C) Cutting the paper into pieces.
VideoChat2 + RRPO: (A) Making the paper look more like a heart.



Question: What caused the sudden fall of the woman in the video?
Options:
(A) She twisted her ankle.
(B) She was tripped.
(C) She was scared by a prank.
(D) Cannot be determined.
LLaVA-Video: (D) Cannot be determined.
LLaVA-Video + DPO: (D) Cannot be determined.
LLaVA-Video + RRPO: (C) She was scared by a prank.

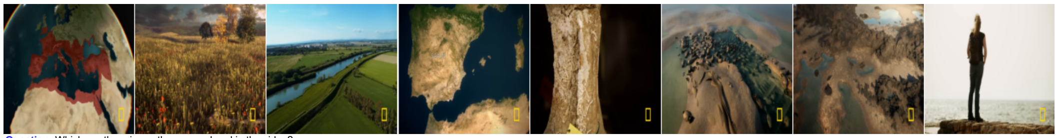


Question: Which of the following options correctly states the number of people visible in the video?
Options:
(A) 5
(B) 4
(C) 6
(D) 3
LLaVA-Video: (D) 3
LLaVA-Video + DPO: (D) 3
LLaVA-Video + RRPO: (B) 4

Question: What is the direction of the camera motion relative to the aircraft carrier?
(A) counterclockwise
(B) downward
(C) upward
(D) clockwise
LongVU: (D) clockwise
LongVU + DPO: (D) clockwise
LongVU + RRPO: (A) counterclockwise

Question: What is the sheep doing?
(A) running
(B) sleeping
(C) eating grass
LongVU: (C) eating grass
LongVU + DPO: (C) eating grass
LongVU + RRPO: (A) running

Figure S6: Qualitative examples on comprehensive video understanding tasks (short).

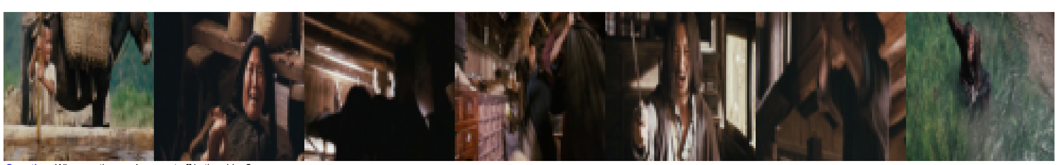


Question: Which are the primary themes explored in the video?
(A) Secrets about Rome found underwater.
(B) How Rome grew into an empire.
(C) The Roman Maritime's prosperity.
(D) The high technologies used to detect sank ships.
VideoChat2: (D) The high technologies used to detect sank ships.
VideoChat2 + DPO: (C) The Roman Maritime's prosperity.
VideoChat2 + RRPO: (A) Secrets about Rome found underwater.

Question: What is this video mainly about?
(A) Two teams complete the mission and survive in the jungle for 24 hours.
(B) Two teams work together to accomplish assigned tasks.
(C) Two teams battled the game planners.
(D) Two squads search for the man in the black suit via VCR.
VideoChat2: (D) Two squads search for the man in the black suit via VCR.
VideoChat2 + DPO: (D) Two squads search for the man in the black suit via VCR.
VideoChat2 + RRPO: (A) Two teams complete the mission and survive in the jungle for 24 hours.

Question: What caused the German athlete in the video to encounter difficulties when attempting to clear the height of 2 meters 27?
Options:
(A) The athlete slipped on the takeoff and lost balance, causing him to miss the jump.
(B) The athlete's coach gave him incorrect instructions, leading to a failed attempt.
(C) The athlete suddenly stopped when he was approaching the crossbar frame, leaving little time for another high jump attempt.
(D) The athlete misjudged the distance and jumped too early, resulting in a failed attempt.
LLaVA-Video: (A) The athlete slipped on the takeoff and lost balance, causing him to miss the jump.
LLaVA-Video + DPO: (A) The athlete slipped on the takeoff and lost balance, causing him to miss the jump.
LLaVA-Video + RRPO: (C) The athlete suddenly stopped when he was approaching the crossbar frame, leaving little time for another high jump attempt.

Question: In which order are the following magics performed in this video?
(a) Finding words in a dictionary.
(b) Drawing a diamond using a match.
(c) Producing a popcoin from eyes.
(d) Seeing things through a bearded dragon.
(e) Inducing judges to choose the cards that they get.
Options:
(A) (c)(a)(e)(d)(b).
(B) (c)(d)(a)(e)(b).
(C) (a)(e)(b)(c)(d).
(D) (a)(b)(d)(c)(e).
LLaVA-Video: (C) (a)(e)(b)(c)(d).
LLaVA-Video + DPO: (C) (a)(e)(b)(c)(d).
LLaVA-Video + RRPO: (B) (c)(d)(a)(e)(b).



Question: Why was the man's ear cut off in the video?
Options:
(A) Because he got into a fight with someone
(B) Another man accidentally cut the wrong person
(C) Because he violated some agreement and was punished
(D) He was retaliated against by others
LongVU: (A) Because he got into a fight with someone
LongVU + DPO: (A) Because he got into a fight with someone
LongVU + RRPO: (B) Another man accidentally cut the wrong person

Question: What genre is this movie?
Options:
(A) Sci-fi
(B) Comedy
(C) Thriller
(D) Romance
LongVU: (A) Sci-fi
LongVU + DPO: (A) Sci-fi
LongVU + RRPO: (B) Comedy

Figure S7: Qualitative examples on comprehensive video understanding tasks (long).

Question: Describe this video.

VideoChat2: The video shows two men standing in the snow, one of them is holding a fishing rod and the other is pulling a sled. The man holding the fishing rod is wearing a gray jacket and a black hat, while the other man is wearing a plaid shirt and a black hat. They seem to be enjoying their time in the snow and are having a good time.

VideoChat2 + DPO: The video shows two men standing in the snow, one of them is holding a fishing rod. They are trying to pull a sled out of the snow. The man holding the fishing rod is wearing a gray jacket and a black hat, while the other man is wearing a plaid shirt and a black hat. They seem to be struggling to pull the sled out of the snow. The snow is deep and white, and there are no other people or objects visible in the video. The sound of the wind can be heard in the background.

VideoChat2 + RRPO: The video shows two men standing in the snow, one of whom is holding a fishing rod. They are standing near a yellow sled that is lying on the ground. One of the men is wearing a hat and a plaid shirt, while the other is wearing a hat and a grey sweater. The man holding the fishing rod is pulling on the line, and the other man is standing nearby. The sled is lying on its side, and there are no other objects visible in the scene. The men seem to be enjoying their time in the snow, and the overall atmosphere of the video is peaceful and serene.

Question: Describe this video.

LongVU: The video captures a sequence of events in a snowy landscape, beginning with a person standing on the snow, holding a long, thin object, possibly a stick or rod, with a small, colorful object attached to it. The person is dressed in winter clothing, including a jacket, pants, and a beanie, suggesting cold weather conditions. The background features leafless trees and a clear sky, indicating a winter setting.

As the video progresses, the person's position changes slightly, and the camera angle shifts to reveal more of the snowy landscape, including a small, colorful object on the ground, which is later identified as a toy airplane. The person's actions suggest they are preparing to launch the airplane, as they are seen holding the rod and the airplane is positioned on the ground, ready for takeoff.

The person then moves closer to the airplane, indicating they are about to launch it. The camera angle changes to provide a clearer view of the person's actions and the airplane's position. The person's stance and the positioning of the airplane suggest they are ready for takeoff.

The video then shows the airplane in motion, with the person holding the rod and the airplane airborne, indicating that it has been launched. The person's posture and the airplane's position suggest they are actively controlling its flight.

Finally, the airplane is seen descending towards the ground, with the person still holding the rod and the airplane in motion. The person's posture and the airplane's position suggest they are actively controlling its descent. The snowy landscape and leafless trees in the background remain consistent throughout the video, providing a continuous winter setting.

LongVU + DPO: The video captures a sequence of events in a snowy landscape, beginning with a person standing on the snow-covered ground, holding a long, thin object, possibly a stick or rod, with a small, colorful object attached to it. The person is dressed in winter clothing, including a jacket, pants, and a beanie, suggesting cold weather conditions. The background features leafless trees and a clear sky, indicating a winter setting.

As the video progresses, the person's position changes slightly, and the camera angle shifts to reveal more of the snowy landscape, including a small, colorful object on the ground, which appears to be a toy or model airplane. The person's posture and the positioning of the object suggest they might be preparing for an activity involving the airplane.

Further into the video, the person is seen holding a remote control device, indicating they are controlling the airplane. The airplane is now visible in the air, suggesting it has been launched or is being flown. The person's actions and the airplane's position imply they are actively engaged in flying the aircraft.

The video continues to show the person adjusting their stance and the remote control device, possibly to maintain control of the airplane. The airplane's position in the air remains consistent, indicating stable flight conditions. The snowy landscape and leafless trees in the background remain unchanged, providing a consistent winter setting throughout the video.

Towards the end of the video, the person is seen holding the remote control device with both hands, possibly preparing for a maneuver or adjusting the airplane's flight path. The airplane is no longer visible in the frame, suggesting it may have moved out of the camera's field of view or is flying at a higher altitude.

In the final scene, the person is standing alone in the snowy landscape, holding the remote control device in one hand, with the other hand raised, possibly signaling or adjusting the flight of the airplane. The person's posture and the positioning of the remote control device suggest they are actively engaged in flying the airplane. The snowy landscape and leafless trees in the background remain consistent, indicating that the video captures a continuous sequence of events in the same winter setting.

LongVU + RRPO: The video captures a sequence of events centered around the operation of a colorful remote-controlled airplane in a snowy outdoor setting. Initially, a person is seen holding a remote control device, presumably controlling the airplane, which is not visible in the frame. The individual is dressed in winter clothing, including a yellow jacket, and is standing on a snowy surface with a backdrop of leafless trees and a clear sky. The ground is marked with tire tracks, suggesting recent activity.

As the video progresses, the person is shown walking towards the camera, indicating movement across the snowy landscape. The remote control device is no longer in view, implying it might have been set aside or is being operated from a distance. The individual's attire and the snowy environment remain consistent throughout these frames.

Subsequently, the focus shifts to the remote-controlled airplane itself, which is now visible on the snow, suggesting it has been recently operated or is being prepared for flight. The individual who was previously controlling the airplane is no longer in the frame, and the camera has zoomed in to capture the airplane's details, including its colorful design and the presence of a propeller.

The video concludes with the airplane in motion, as indicated by the blurred propeller and the disturbed snow around it. This suggests that the airplane has been activated and is either taking off or landing. The camera maintains its focus on the airplane, capturing the dynamic action against the unchanged snowy backdrop and leafless trees, emphasizing the contrast between the moving airplane and the stillness of the surrounding environment.

Question: Describe this video.

LLaVA-Video: The video depicts a winter scene where two individuals are engaged in an activity involving a small, yellow model airplane. The setting is an open, snow-covered field with a backdrop of leafless trees, indicating a cold season. The ground is covered in a thick layer of snow, and the sky appears overcast.

Objects and Individuals:
1. **Two Individuals:**
- The first individual is wearing a plaid jacket, green pants, and a black cap. He is holding a remote control device, likely used to operate the model airplane.
- The second individual is dressed in a gray jacket, dark pants, and a beanie. He is handling the model airplane, which is predominantly yellow with some blue and red accents.

2. **Model Airplane:**
- The model airplane is small, yellow, and has a simple design with a single set of wings and a tail. It appears to be a glider or a basic remote-controlled plane.

Activity Description:
- The person in the gray jacket is seen adjusting or preparing the model airplane on the snow. He seems to be ensuring that the airplane is stable and ready for operation.
- The person in the plaid jacket is holding the remote control, suggesting he is about to operate the airplane.
- As the video progresses, the person in the gray jacket steps back slightly, allowing the person in the plaid jacket to take control of the remote and start operating the model airplane.
- The model airplane is then seen being maneuvered by the person in the plaid jacket, who is using the remote control to steer it across the snowy field.

Background and Environment:
- The background consists of a vast expanse of snow with tracks from previous activities, possibly from other model airplanes or vehicles.
- The trees in the distance are bare, reinforcing the winter setting.

Conclusion:
The video captures a moment of leisure and enjoyment as two individuals engage in flying a model airplane in a snowy field. The preparation and coordination involved in this activity highlight the teamwork and attention to detail required for such hobbies. The serene winter landscape adds to the overall ambiance of the scene.

LLaVA-Video + DPO: The video depicts a winter scene where two individuals are engaged in an activity involving a small, yellow model airplane. The setting is an open, snow-covered field with a backdrop of leafless trees, indicating a cold season.

Objects and Individuals:
1. **Two Individuals:**
- The first individual is wearing a plaid jacket, green pants, and a black cap. He is holding a remote control device, likely used to operate the model airplane.
- The second individual is dressed in a gray jacket, dark pants, and a beanie. He is handling the model airplane, adjusting its position on the snow.

2. **Model Airplane:**
- The model airplane is predominantly yellow with some blue and red accents. It has a single propeller at the front and appears to be a simple, single-engine design.

Activity Description:
- The person in the plaid jacket is operating the remote control, which suggests he is controlling the movement of the model airplane.
- The person in the gray jacket is adjusting the position of the model airplane on the snow, possibly preparing it for flight or ensuring it is stable.
- Throughout the video, the person in the gray jacket moves around the model airplane, making adjustments while the person in the plaid jacket maintains control over the remote.

Background and Environment:
- The background consists of a vast, snow-covered field with tracks from previous activities, indicating that this area is frequently used for similar activities.
- The trees in the distance are bare, suggesting it is either late autumn or winter.

Conclusion:
The video captures a moment of preparation and coordination between two individuals as they work together to operate a model airplane in a snowy environment. The detailed actions and attire suggest a recreational activity that requires both technical skill and teamwork.

LLaVA-Video + RRPO: The video depicts a winter scene where two individuals are engaged in an activity involving a small, yellow model airplane. The setting is an open, snow-covered field with a backdrop of leafless trees, indicating a cold season.

Objects and Individuals:
1. **Two Individuals:**
- The first individual is wearing a plaid jacket, green pants, and a black cap. He is holding a remote control device, likely used to operate the model airplane.
- The second individual is dressed in a gray jacket, dark pants, and a beanie. He is handling the model airplane, adjusting its position on the snow.

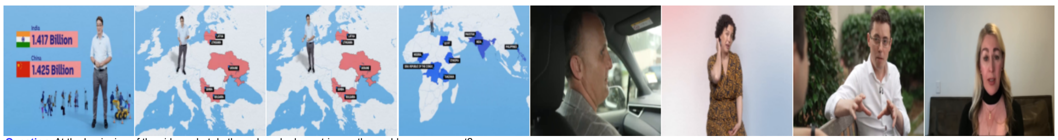
2. **Model Airplane:**
- The model airplane is predominantly yellow with some blue and red accents. It has a single propeller at the front and appears to be a simple, single-engine design.

Activity Description:
- The first individual is standing and holding a remote control device, which he uses to operate the model airplane. His stance suggests he is focused on controlling the plane.
- The second individual is crouched down next to the model airplane, making adjustments to its position on the snow. He seems to be ensuring that the airplane is stable and ready for operation.
- Throughout the video, the second individual moves around the airplane, occasionally stepping back to assess his work before making further adjustments.
- The first individual remains stationary, concentrating on the remote control, while the second individual moves around the airplane, making sure it is properly positioned.

Contextual Analysis:
- The activity involves both individuals working together to prepare the model airplane for flight. The person with the remote control is responsible for operating the plane once it takes off, while the other ensures the plane is correctly set up on the ground.
- The snowy environment provides a suitable surface for the model airplane to glide smoothly without the risk of sinking into softer ground.
- The presence of multiple tracks in the snow suggests that this area is frequently used for similar activities or other forms of recreation.

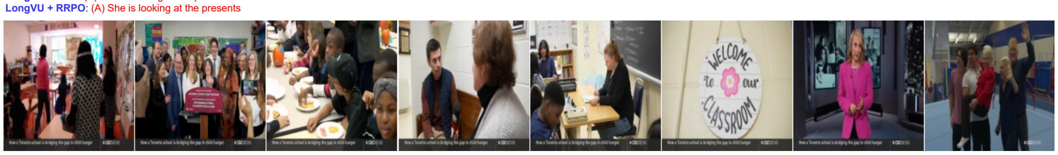
Conclusion:
The video captures a moment of teamwork and preparation as the two individuals collaborate to get the model airplane ready for flight. The snowy landscape and the individuals' attire indicate a cold, possibly windy day, which is ideal for flying model airplanes. The focus on the model airplane and the individuals' actions highlights their dedication to ensuring the plane is properly set up for a successful flight.

Figure S8: Building on the capabilities of the base models, we observe that RRPO-aligned models may still make mistakes or exhibit hallucinations in detailed video description tasks. For instance, VideoChat2 continues to display similar hallucinations after both DPO and RRPO training, as seen in the base model. In contrast, for LongVU, while both the base and DPO-aligned models hallucinate, the RRPO-aligned variant avoids such errors. Finally, in the case of LLaVA-Video, the RRPO-aligned model retains the base model's reliable behavior, as neither exhibits hallucinations.



Question: At the beginning of the video, what do the red-marked countries on the world map represent?
(A) The population of these countries is declining.
(B) The population of these countries is growing.
(C) These countries are experiencing famine.
(D) These countries are experiencing war.
VideoChat2: (B) The population of these countries is growing.
VideoChat2 + DPO: (B) The population of these countries is growing.
VideoChat2 + RRPO: (B) The population of these countries is growing.

Question: What is the young beautiful woman looking at while preparing for the new year?
Options:
(A) She is looking at the presents
(B) She is looking at the golden christmas toy
(C) She is looking at the christmas lights
(D) She is looking at the christmas stockings
LongVU: (A) She is looking at the presents
LongVU + DPO: (A) She is looking at the presents
LongVU + RRPO: (A) She is looking at the presents



Question: At the end of the video, what is the sentiment of the four Russian defectors towards a woman named Judy Tulk?
Options:
(A) Gratitude.
(B) Missing.
(C) Admiration.
(D) Cannot be determined.
LLaVA-Video: (D) Cannot be determined.
LLaVA-Video + DPO: (D) Cannot be determined.
LLaVA-Video + RRPO: (D) Cannot be determined.

Figure S9: We also observe that RRPO-aligned models may still exhibit limitations in long video understanding tasks, primarily due to architectural constraints of the base model in processing extended frame sequences, as well as computational constraints during RRPO training that limit the use of long video inputs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions of this paper are mentioned in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We present the mathematical derivation of our theoretical claims.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental results are fully reproducible, and sufficient details are shared in the paper. Moreover, we release the code and data through an anonymized repository during the review process for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code are shared through an anonymized repository during the review process for reproducibility.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The necessary details for the experimental settings are shared in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We test for statistical significance and consider performance improvements from the fine-tuned model over the base model statistically significant at the 95% confidence level, based on Standard Error.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details of the computation requirements are discussed in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader impact of this work is discussed in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: This work involves finetuning publicly available, off-the-shelf large video-language models (LVLMs) to enhance their performance across diverse video understanding tasks. The base LVLMs were originally trained on carefully filtered public datasets and are already widely adopted in the research community. Moreover, our finetuning process uses open-source, carefully curated training data. We do not anticipate this process introducing any new safety concerns.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Licenses for existing assets used in this work are mentioned in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets`

has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we introduce code, model, and data in this paper, which are available via an anonymous link during the review period and will be publicly released afterward. We also provide sufficient details of dataset/code/model as part of our submission throughout the main paper and the appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This work does not involve the use of LLMs in any way that affects the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.