Oracle-RLAIF: An Improved Fine-Tuning Framework for Multi-modal Video Models through Reinforcement Learning from Ranking Feedback

Derek Shi[1,2] **Ruben Glatt**[1] **Christine Klymko**[1] **Shubham Mohole**[1,3]

Hongjun Choi[1] **Shashank Kushwaha**[1] **Sam Sakla**[1] **Felipe Leno da Silva**[1]

[1] Lawrence Livermore National Laboratory, CA, USA
[2] University of California Los Angeles, CA, USA
[3] Cornell University, NY, USA
{glatt1,klymko1,mohole1,kushwaha1,choi22,sakla1,leno}@llnl.gov
derekshi@ucla.edu

Abstract

Recent advances in large video-language models (VLMs) rely on extensive fine-tuning techniques that strengthen alignment between textual and visual comprehension. Leading pipelines typically pair supervised fine-tuning (SFT) with reinforcement learning from preference data to enhance video comprehension. However, as VLMs scale in parameter size, so does the cost of gathering enough human feedback. To make fine-tuning more cost-effective, recent frameworks explore reinforcement learning with AI feedback (RLAIF), which replace human preference with AI "as a judge". Current RLAIF frameworks rely on a specialized reward model trained with video narratives to create calibrated scalar rewards– an expensive and restrictive pipeline. We propose Oracle-RLAIF, a novel framework that replaces the trained reward model with a more general Oracle ranker which acts as a drop-in model ranking candidate model responses rather than scoring them. Alongside Oracle-RLAIF, we introduce $GRPO_{rank}$, a novel rank-based loss function based on Group Relative Policy Optimization (GRPO) that directly optimizes ordinal feedback with rank-aware advantages. Empirically, we demonstrate that Oracle-RLAIF consistently outperforms leading VLMs using existing fine-tuning methods when evaluated across various video comprehension benchmarks. Oracle-RLAIF paves the path to creating flexible and data-efficient frameworks for aligning large multi-modal video models with reinforcement learning from rank rather than score.

1 Introduction

Current multi-modal foundation models are capable of providing fluent image captioning, accurate video question answering, and video reasoning capabilities [Ahn et al., 2024, Liu et al., 2023, Li et al., 2023b]. Typically, state-of-the-art (SOTA) multi-modal models are trained following a two-step process. First, supervised fine-tuning (SFT) through human-annotated videos allows video language models (VLMs) to produce syntactically correct and relevant answers [Luo et al., 2023]. Next, following SFT, a reinforcement-learning-from-human-feedback (RLHF) [Ouyang et al., 2022] phase uses human preference over possible model outputs to further video comprehension in the

Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

VLM. Although human labeled responses provide quality ground-truth data, this method introduces substantial bottlenecks of labeling inefficiency and cost. As a result, a few works in the literature started to rely on AI judge models to replace human feedback– resulting in reinforcement-learning-from-AI-feedback (RLAIF) [Ahn et al., 2024]. RLAIF consists of using a model capable of outputting a "reward" signal for an arbitrary prompt answer such that it can be used for fine-tuning. However, for many tasks, building a model capable of generating consistent and grounded rewards for any arbitrary combination of prompts and outputs has proven challenging [Shen et al., 2024, Chen et al., 2024].

We address these limitations by proposing a more flexible RLAIF framework called Oracle-RLAIF. Instead of relying on a fully functioning reward model capable of scoring any combination of prompts and responses, we only require an *Oracle model* capable of ranking responses in order of quality. This makes our framework broadly applicable to a variety of scenarios, including fine-tuning with feedback from a general-purpose closed-source model, distilling knowledge from a large model into a smaller one, or leveraging a legacy system based on a different AI paradigm. Moreover, we introduce a novel GRPO extension we call $GRPO_{rank}$, which directly processes ranks from Oracle to effectively fine-tune the initial multi-modal model. Across multiple video evaluation datasets, our Oracle-RLAIF model directly outperforms previous state-of-the-art fine-tuned VLMs. Our primary contributions are summarized as follows:

- We introduce Oracle-RLAIF, a novel rank-based RLAIF framework, utilizing a drop-in Oracle ranker, relaxing the need of a fully-functional reward model.
- We develop a robust GRPO modification, which effectively guides learning through direct rank signal during reinforcement learning.
- We extensively validate our proposed rank-based framework by training and evaluating our Oracle-RLAIF VLM. When directly compared to current fine-tuning techniques, Oracle-RLAIF improves video comprehension performance across benchmark datasets.

2 Background

We begin by providing all necessary background information for multi-modal learning as well as fine-tuning techniques for video language models. Additionally, we include an in-depth description of the framework we build off and the specific limitations we address.

2.1 Multi-Modal Learning

Multi-modal learning trains models that can comprehend inputs from various modalities such as vision, audio, and language [Radford et al., 2021, Alayrac et al., 2022, Li et al., 2023a]. In video language models (VLMs), this means generating textual responses conditioned on visual information. These models typically comprise a vision encoder, which converts raw images into feature representations aligned with text [Radford et al., 2021, Li et al., 2023a], and a large causal language model (e.g., LLaMA, GPT [OpenAI, 2023, Touvron et al., 2023]) that generates responses based on those features.

To enhance video understanding, models are further trained via supervised fine-tuning (SFT) on video–question–answer triplets, grounding generation in visual reasoning tasks. However, extensive spatial and temporal understanding is still limited. Therfore, to further align models, reinforcement learning after SFT has become necessary to achieve SOTA video question answering performance [Liu et al., 2023, Ahn et al., 2024].

2.2 Reinforcement Learning techniques for Fine-Tuning Video Language Models

Two main frameworks have previously been implemented to conduct reinforcement learning, as an evolution of the long literature in leveraging human feedback for learning [Silva and Costa, 2019]. Reinforcement Learning from Human Feedback (RLHF) is the process of using human feedback to choose a preferred response from two distinct model outputs for a given query (Figure 1). Due to the high cost of human labeling, researchers have introduced RLAIF [Bai et al., 2022, Lee et al., 2023, Su et al., 2023] where a *context aware AI* replaces the human as the judge (Figure 1). Recently, VLM-RLAIF [Ahn et al., 2024] applied this to multi-modal models by scoring two candidate responses from an initial supervised fine-tuned VLM (VLM-SFT) using a context aware AI judge. VLM-RLAIF validated their approach presenting significantly improved video question-answering abilities

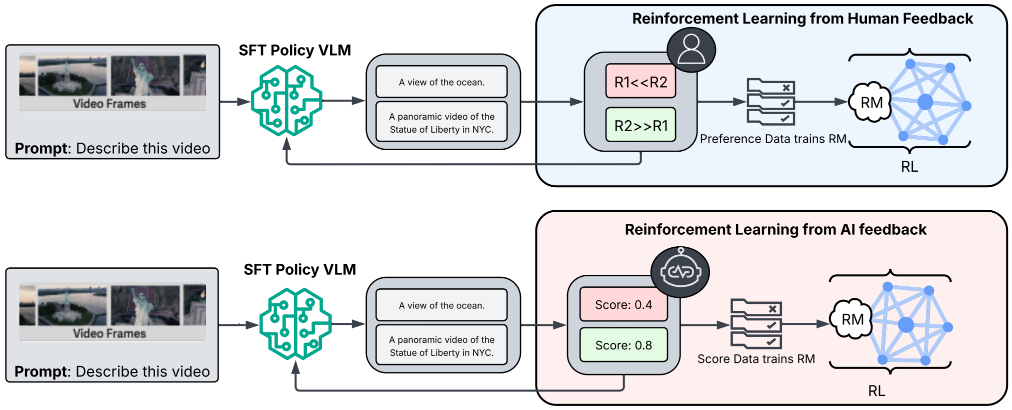

Figure 1: Dataset creation pipelines for RLHF and RLAIF, illustrating how preferences are used to train reward models which are then used in the reinforcement learning loop to align the initial VLM.

when compared to their initial VLM before RLAIF. Additionally, VLM-RLAIF currently achieves state-of-the-art performance in video comprehension [Ahn et al., 2024].

For both human and AI feedback, the purpose of the preference data is to train a reward model to score answers. Subsequently, these rewards are used to guide an initial VLM towards preferred answers [1] [Hong et al., 2025]. When in possession of this reward model, any canonical RL algorithm can be used to fine-tune the model responses. In the LLM fine-tuning space, specific algorithms are preferred since they have been demonstrated to be effective.

Proximal Policy Optimization (PPO) [Schulman et al., 2017] is arguably the original algorithm for LLM fine-tuning, and notably also the algorithm employed by the original VLM-RLAIF framework. PPO consists of optimizing the model towards improving reward while constraining the maximum deviation from the original policy. The loss function for PPO is:

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\!\big(r_t(\theta)\,\hat{A}_t,\ \text{clip}\big(r_t(\theta), 1-\varepsilon, 1+\varepsilon\big)\,\hat{A}_t\big)\ -\ \beta\, D_{\text{KL}}\big[\,\pi_{\theta_{\text{old}}}\,\|\,\pi_\theta\big]\Big] \tag{1}$$

For a trajectory $\big(s_t, a_t\big)$ with return $\hat{R}_t$ and baseline value $V_\phi(s_t)$,

$$\hat{A}_t \;=\; \hat{R}_t - V_\phi(s_t), \qquad r_t(\theta) \;=\; \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)} \tag{2}$$

The expected reward is represented through the advantage function $\hat{A}_t$, encoding how much better than the "average" action that particular action is expected to be. PPO is used in the VLM-RLAIF framework to guide the initial SFT model based on rewards from the trained RM. Importantly, unlike the original RLAIF framework, we only have the rankings instead of a score-based reward model, thus PPO is not optimal for our framework.

Group Relative Policy Optimization (GRPO) [Shao et al., 2024] addresses some limitations of PPO including unstable updates from reward magnitudes and the requirement of a trained value function to compute advantage. As introduced by DeepSeekMath, GRPO extends PPO to sample relative performance over $\boldsymbol{G_i}$ candidate responses for one query $\boldsymbol{q}$.

[1] Some RLHF techniques do not require an explicit reward model in memory, however they follow assumptions regarding the preference distribution, thus the same general problem stands. [Rafailov et al., 2023]

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left(r_t(\theta)\hat{\boldsymbol{A}}_{\boldsymbol{i,t}},\ \text{clip}\left(r_t(\theta), 1-\epsilon,\ 1+\epsilon\right)\hat{\boldsymbol{A}}_{\boldsymbol{i,t}}\right)\right\}\right.$$
$$\left. - \beta \cdot D_{\text{KL}}\left[\pi_{\theta_{\text{old}}}(\cdot \mid \boldsymbol{q}) \,\|\, \pi_\theta(\cdot \mid \boldsymbol{q})\right] + c_{\text{entropy}} \cdot \mathcal{H}[\pi_\theta(\cdot \mid \boldsymbol{q})]\right] \tag{3}$$

where the policy importance ratio is defined as:

$$r_t(\theta) = \frac{\pi_\theta(o_{i,t} \mid \boldsymbol{q}, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \boldsymbol{q}, o_{i,<t})} \tag{4}$$

and the advantage uses **normalized reward**

$$\hat{\boldsymbol{A}}_{\boldsymbol{i,t}} = \frac{R(\boldsymbol{q}, o_i) - \mu_R(\boldsymbol{q})}{\sigma_R(\boldsymbol{q})} \tag{5}$$

where:

$\pi_\theta(o_{i,t} \mid \boldsymbol{q}, o_{i,<t})$ denotes the probability of token $o_{i,t}$ under the current policy π_θ,
$\pi_{\theta_{\text{old}}}(o_{i,t} \mid \boldsymbol{q}, o_{i,<t})$ denotes the same probability under the frozen (old) policy $\pi_{\theta_{\text{old}}}$.

Since rewards are rescaled per-query, GRPO is robust to varying reward magnitudes and eliminates the need for a separate value function (Eq 5). Used by DeepSeek to finetune their R1 reasoning models, GRPO has already been proven effective for LLM post-training alignment [DeepSeek-AI, 2025]. However, GRPO's chief advantage over PPO comes from gaining relational performance from multiple candidate responses, we extend this idea further by introducing $GRPO_{rank}$ (Sec 4) which replaces reward score with direct rank.

3 Reinforcement Learning from AI Feedback through Oracle Preferences

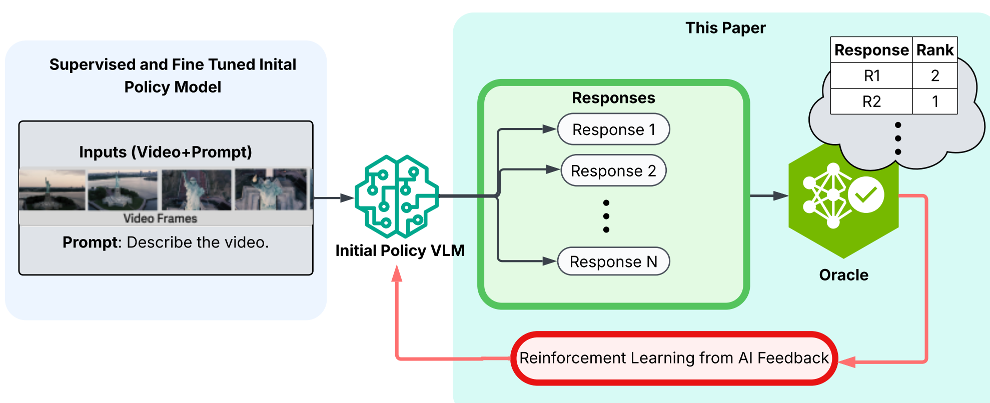

Figure 2: The general pipeline of fine tuning the initial policy VLM using Oracle-RLAIF. The light blue shading on the left indicates SFT training already applied to the initial policy VLM. The cyan shading on the right represents fine-tuning using our RLAIF Oracle ranker pipeline. Where the policy outputs *N responses* ranked by Oracle to fine-tune our VLM.

We propose the Oracle-RLAIF fine-tuning framework where a drop-in Oracle ranker replaces the trained RM in previous frameworks. Specifically, where VLM-RLAIF relied on their supervised fine-tuned VLM to label videos with additional narrative context to train their reward model, Oracle-RLAIF uses a stand alone Oracle ranker which only needs to be capable of ordering response quality.

This eliminates the need to calibrate for trustworthy score magnitude and creates a much more flexible fine-tuning framework. Our work does not change the process of creating the initial SFT model, but rather alters the RLAIF loop to take advantage of rank-based learning as detailed below.

Oracle-RLAIF begins from a pretrained SFT video language model that is able to accurately generate responses from visual and textual input. Starting with this model, we enter the RLAIF iterative process where the initial SFT model generates N responses from a single prompt. The multi-modal Oracle ranker is then prompted to rank the candidate responses based on quality and relevance to the visual input. Using these ranks, our policy optimization algorithm finetunes the initial SFT model to align responses according to the Oracle rankings. The Oracle-RLAIF framework does not require a trained reward model or a value model to calculate expected reward as in VLM-RLAIF [Ahn et al., 2024] or other similar frameworks. Our suggested framework works purely off the initial policy model and the Oracle ranker.

4 Fine-tuning Policy Models using Oracle Ranking

In order to provide learning signal from Oracle-ranked feedback, we create $GRPO_{rank}$. While building upon the objective function structure from GRPO [Shao et al., 2024], we significantly modify the calculation of the advantage function. These modifications take into account several key intuitions when replacing reward score with direct rank. We expand on these intuitions in the next subsection.

4.1 Rank Adapted GRPO Objective

We introduce $GRPO_{rank}$ to optimize the policy within our Oracle-RLAIF framework. It applies non-linear penalization of rank errors, assigning larger penalties and smaller advantages when the predicted rank deviates more from the Oracle ground truth. It also penalizes false promotion of low-quality responses, encouraging the model to both prioritize high-quality outputs and suppress poor ones. To capture these intuitions, we use a normalized Discounted Cumulative Gain (nDCG) penalty that compares the *predicted rank* (based on the model's internal log probabilities) to the *ground-truth rank* assessed by the Oracle ranker. This penalty captures both position sensitivity and error severity, and is used to construct the advantage function for our policy gradient updates.

For stability, $GRPO_{rank}$ generates on-policy responses while using a frozen reference policy for importance sampling. Our loss function modifies the original GRPO loss(Eq 3) by replacing the advantage term with $\hat{A}_{\text{rank}}$:

$$\mathcal{L}_{GRPO_{rank}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\Big(r_t(\theta)\hat{A}_{rank},\ \text{clip}\left(r_t(\theta), 1-\epsilon,\ 1+\epsilon\right)\hat{A}_{rank}\Big)\right\}\right.$$
$$\left. - \beta \cdot D_{\text{KL}}\left[\pi_{\theta_{\text{old}}}(\cdot \mid \boldsymbol{q}) \,\|\, \pi_\theta(\cdot \mid \boldsymbol{q})\right] + c_{\text{entropy}} \cdot \mathcal{H}[\pi_\theta(\cdot \mid \boldsymbol{q})]\right] \tag{6}$$

The advantage function $\hat{A}_{\text{rank}}$ in $GRPO_{rank}$ measures how much better each response i is when compared to the model's "average response" and is defined as:

$$\hat{A}_{\text{rank}} = \mathbb{E}_{j\in G_i}[\delta_j] - \delta_i \tag{7}$$

with the expected group penalty for K responses within group G_i:

$$\mathbb{E}_{j\in G_i}[\delta_j] = \frac{1}{K}\sum_{j\in G_i}\delta_j \tag{8}$$

and the penalty term δ_i represents the deviation of the model's predicted rankings from the Oracle's ranking. This penalty is calculated using normalized Discounted Cumulative Gain (nDCG):

$$\delta_i = 1 - \text{nDCG}_i = 1 - \frac{\text{DCG}(\hat{\text{rank}}_i)}{\text{DCG}(\text{rank}_i)} \tag{9}$$

where $\hat{\text{rank}}_i \in \{0, \ldots, K-1\}$ is the model's predicted ranking computed from ordering the policy model's internal log probabilities and $\text{rank}_i \in \{0, \ldots, K-1\}$ is from the Oracle ranker. Finally, DCG is computed as:

$$\text{DCG(rank)} = \frac{\frac{1}{1+\text{rank}}}{\log_2(2 + \text{rank})} \tag{10}$$

Algorithm 1 Iterative $GRPO_{rank}$

Require: Initial policy model θ_{init}; Oracle model $\mathcal{O}_\phi$; task prompts $\mathcal{D}$; hyperparameters ϵ, β

Initialize policy model: $\pi_\theta \leftarrow \theta_{\text{init}}$
for epoch = $1, \ldots, E$ **do**
    Store reference policy: $\pi_{\text{ref}} \leftarrow \pi_\theta$
    for training step **do**
        Sample batch $\mathcal{D}_b \subset \mathcal{D}$
        Generate G responses $\{o_i\}_{i=1}^{G}$ from $\pi_\theta(\cdot \mid q)$ for each $q \in \mathcal{D}_b$
        $rank_i \leftarrow \mathcal{O}_\phi(o_i), \forall i \in \{1, \ldots, G\}$
        Compute log-probabilities of responses under current policy:

$$\ell_i = \log \pi_\theta(o_i \mid q) = \sum_{t=1}^{|o_i|} \log \pi_\theta(o_{i,t} \mid o_{i,<t}, q) \quad \forall i \in \{1, \ldots, G\}$$

        Compute predicted ranks $\hat{\mathbf{rank}}$ by sorting log-probabilities $\boldsymbol{\ell}$ in descending order:

$$\hat{\mathbf{rank}} = \text{argsort}_{\text{desc}}(\boldsymbol{\ell})$$

        Update π_θ via gradient descent according to $\mathcal{L}_{\text{GRPO}_{\text{rank}}}$ (Eq 6)
    end for
end for
return final policy π_θ

We detail the iterative $GRPO_{rank}$ fine-tuning algorithm in Algorithm 1. At each epoch, batches of prompts are sampled, and candidate responses are generated using the current policy model. The *Oracle model* provides relative rankings over responses in each group, from which nDCG-based penalties are computed. These penalties are then used to define an advantage function $\hat{A}_{\text{rank}}$, which captures how much better or worse each response is compared to the average performance in its group. The policy is then updated by maximizing a clipped surrogate objective (Eq. 6), incorporating KL and entropy regularization to ensure stability. This allows the model to iteratively prefer responses ranked higher by the oracle while preventing large policy deviations.

Mathematical Properties

To further understand the motivation behind the elements in our $GRPO_{rank}$ objective, we highlight key mathematical properties that guided our formulation. This section explains the desired properties and how our formulation satisfies them. Additional examples from Appendix A.2 illustrate advantages via hypothetical rank configurations.

1. Zero-Sum Property Within Groups. Our advantage formulation:

$$\hat{A}_{\text{rank}} = \mathbb{E}_{j \in G_i}[\delta_j] - \delta_i$$

ensures that total advantage across group members sums to zero, enforcing relative comparison within each group:

$$\sum_{i \in G} \hat{A}_{i,t} = \sum_{i \in G} (\mathbb{E}_{j \in G}[\delta_j] - \delta_i) = K \cdot \mathbb{E}_{j \in G}[\delta_j] - \sum_{i \in G} \delta_i = K \cdot \left(\frac{1}{K} \sum_{j \in G} \delta_j \right) - \sum_{i \in G} \delta_i = 0$$

2. Boundedness of Penalty. Since $\text{rank}_i, \hat{\text{rank}}_i \in \{0, \ldots, K-1\}$, DCG values fall in:

$$\text{DCG}_i \in \left[\frac{1}{(1+K-1)\cdot\log_2(K+1)},\, 1\right]$$

Thus the normalized DCG, $\text{nDCG}_i = \frac{\text{DCG}_i(\hat{\text{rank}}_i)}{\text{DCG}_i(\text{rank}_i)} \in (0,1]$, leading to bounded penalties:

$$\delta_i = 1 - \text{nDCG}_i \in [0,1)$$

ensuring numerically stable and interpretable advantage values.

3. Position-Sensitive Discounting. The use of logarithmic discounting in:

$$\text{DCG(rank)} = \frac{\frac{1}{1+\text{rank}}}{\log_2(2+\text{rank})} \tag{11}$$

ensures that rank errors at the top are penalized exponentially more than those at the bottom. This encodes our preference to rank correctly the top answers, which are the ones that will be displayed to the user, while the relative rankings of worse answers are not as important. For example:

True Rank 0, Predicted Rank 1: $\text{nDCG} = 0.7925 \Rightarrow \delta = 0.2075$
True Rank 4, Predicted Rank 3: $\text{nDCG} = 0.9757 \Rightarrow \delta = 0.0243$

Finally, the highest penalty originates from a low-ranked high-quality response:

True Rank 0, Predicted Rank 4: $\text{nDCG} = 0.5170 \Rightarrow \delta = 0.4830$

5 Empirical Evaluation

Our primary goal in our experimentation is to assess whether Oracle-RLAIF improves upon leading fine-tuning frameworks for VLMs– specifically the current SOTA VLM-RLAIF. We benchmark using datasets which test a model's capacity to interpret and describe visual events. To compare Oracle-RLAIF against VLM-RLAIF, we apply both frameworks to the same initial SFT policy model in Figure 2). Additionally, we compare against reported results of other strong baselines to comprehensively assess the contribution of our rank-based policy optimization approach.

5.1 Experimental Setup

Evaluation Protocol: Since our main purpose is to evaluate the fine-tuning performance of our proposal compared to VLM-RLAIF, we start both approaches with the same SFT model (the VLM-SFT 7B checkpoint as published in Ahn et al. [2024]). From there, VLM-RLAIF is trained as described in their paper (Appendix A.1) with the released public implementation– with only a few training configurations modified below. We also use their publicly released and pre-trained reward model to train VLM-RLAIF (7B) for 4 epochs and a rollout batch size of 64 for more frequent gradient updates (modified from 1 epoch and 256 batch size in the original publication to allow for more policy updates)

In order to ensure performance gains are due to better fine-tuning strategy rather than feedback performance, we train our Oracle ranker reward model in the same pipeline as the VLM-RLAIF reward model. However, unlike VLM-RLAIF, we omit caption data to train Oracle ranker, simulating a drop-in reward model not specifically tuned with video narratives. We use the same training configurations from VLM-RLAIF, and set the number of candidate responses to 5 when finetuning Oracle-RLAIF (7B).

Both models are trained using 4×NVIDIA H100 80GB GPUs with *Quantized Low-Rank Adapter (QLoRA)* [Dettmers et al., 2023], which enables efficient fine-tuning by combining 4-bit quantization with low-rank adapters. The fully trained models are then evaluated as described in the next subsections.

Evaluation Datasets and Benchmarks We compared the models across two distinct evaluation regimes. In the first, we follow the evaluation pipeline used in VLM-RLAIF by benchmarking across MSVD, MSRVTT, and ActivityNet [Wu et al., 2017, Xu et al., 2016, Yu et al., 2019]. Three video question answering datasets which target action recognition, temporal reasoning, and overall multi-modal understanding. Since the benchmark uses open-ended questions, the model's responses must be evaluated by an LLM (in this case GPT-3.5-turbo) in comparison to human-annotated ground-truth labels. This follows the methodology described in Li et al. [2023c]. The LLM as a judge considers five key dimensions (response relevance to video content, detail capture, contextual understanding, temporal reasoning, and consistency) to create a numeric score from 0 to 5 as well as a binary "yes" or "no" indicating correct response.

In the second evaluation regime, we instead benchmark against the better performing "final" checkpoint of the VLM-RLAIF 7B model as trained and evaluated *by the authors*.[2] In this comparison, we take advantage of a more contemporary dataset Video-MME [Fu et al., 2025] which was not available at the time of VLM-RLAIF's publication. By design, Video-MME does not originate from existing video evaluation datasets, allowing no possibility of data leakage into training. Additionally, unlike prior benchmarks which rely on GPT-judged open-ended answers, Video-MME establishes objective and reproducible evaluation through multiple choice question answering. Thus, we are able to directly compare accuracy across categories without relying on language model judgements. Video-MME is currently widely used for benchmarking, and we consider this our most meaningful experiment.

5.2 Results

Table 1: Performance of Oracle-RLAIF versus baseline VLMs in zero-shot question answering across MSVD, MSRVTT, and ActivityNet.

Model	**Fine-tuning method**	**MSVD-QA**		**MSRVTT-QA**		**ActivityNet-QA**	
		Acc.	Score	Acc.	Score	Acc.	Score
VideoLLaMA ([Zhang et al., 2024])	SFT	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT ([Maaz et al., 2023])	SFT	64.9	3.3	49.3	2.9	35.2	2.7
BT-Adapter ([Liu et al., 2024])	SFT	67.5	3.7	57.0	3.2	45.7	3.2
Video-LLaVA ([Lin et al., 2024])	SFT	70.7	3.9	59.2	3.5	45.3	3.3
LLaMA-VID ([Li et al., 2024])	SFT	69.7	3.7	57.7	3.2	47.4	3.23
VideoChat2 ([Li et al., 2023b])	SFT	70.0	3.9	54.1	3.3	49.1	3.3
VLM-SFT (baseline)	SFT	67.2	3.6	52.4	3.0	44.1	3.2
VLM-RLAIF	RLAIF	68.5	3.6	54.2	3.1	46.1	3.4
Oracle-RLAIF (ours)	RLAIF	**72.9**	**3.9**	**59.2**	**3.7**	**48.1**	**3.5**
Δ (Oracle-RLAIF − VLM-RLAIF)		**+4.4%**	**+0.3**	**+5.0%**	**+0.6**	**+2.0%**	**+0.1**

Detailed Analysis Table 1 shows the results for our first experiment.[3] Oracle-RLAIF outperforms all baselines we evaluated in video-question answering performance. Specifically, we observe consistent gains across all three benchmarks when compared to our reproduced VLM-RLAIF model following the author's framework. This demonstrates that our rank-based optimization approach yields more effective policy updates compared to VLM-RLAIF, which employs scalar reward modeling and PPO updates. The original VLM-RLAIF publication reports significantly better results than we could achieve on ActivityNet— which we attribute to the use of ActivityNet caption data in their reward model training which we omitted, thus creating unfair evaluation within this dataset.

Therefore, in our second experiment reported in Table 2, we use Video-MME ensuring no data leakage from training to evaluation. Here, Oracle-RLAIF significantly outperforms the original VLM-RLAIF checkpoint, demonstrating superior generalization and robustness. Overall, Oracle-RLAIF achieves

[2] Despite our best attempt to replicate the results from the original paper using the public implementation, a few key scripts were missing causing slightly lower performances than the final published checkpoint. In the second evaluation, however, we use the best performing model, even though we could not replicate its training.

[3] We do not replicate experiments in all of the SFT models, and instead report their published results from prior works.

Table 2: Video-MME: performance comparison across short and medium videos

Category	Oracle-RLAIF	VLM-RLAIF	Δ
Temporal Perception	**71.0%**	49.8%	+21.2
Action Recognition	**39.0%**	27.3%	+11.7
Object Reasoning	**49.5%**	38.3%	+11.2
Object Recognition	**46.0%**	35.2%	+10.8
Counting Problem	**27.1%**	19.4%	+7.6
OCR Problems	**38.1%**	34.6%	+3.5
Attribute Perception	**42.2%**	39.4%	+2.8
Temporal Reasoning	**29.9%**	27.7%	+2.2
Action Reasoning	**33.5%**	31.8%	+1.7
Information Synopsis	48.5%	**49.4%**	-0.9
Spatial Perception	32.5%	**35.1%**	-2.6
Spatial Reasoning	51.2%	**55.0%**	-3.8
Overall Accuracy	**42.4%**	36.2%	+6.2

a +6.2% percent improvement in average accuracy, with significant improvements in *Temporal Perception* (+21.2%), *Action Recognition* (+11.7%), and *Object Reasoning* (+11.2%)– tasks which involve physical actions and causal events. For instance, Oracle-RLAIF accurately answers questions such as "*Which action is not included in the third magic trick?*" from *Action Recognition*, and successfully understands time-based cues in questions like "*In which part of the video is the woman in the blue top interviewed?*" from *Temporal Perception*. These improvements results point to the core strength of our rank-based optimization to, which improves model alignment with temporally and causally grounded responses. These results point to the core strength of our rank-based optimization to improve model alignment with temporally and causally grounded responses.

In contrast, Oracle-RLAIF shows performance declines in *Spatial Perception*, *Spatial Reasoning*, and *Information Synopsis*. We hypothesize that these categories contain higher ambiguity or abstraction, causing rank-based optimization to be less effective. For example, *Spatial Reasoning* questions such as "*What holiday was the video most likely recorded during?*" or "*What can be inferred about the setting from the lighting and background?*" rely on implicit cues that are not effectively optimized through relative ranking. Additionally, *Spatial Perception* questions: "*Which option correctly indicates where the "psychological tip" is written in white on the board when illustrating the third suggestion in the video?*" require spatial awareness across frames, which may benefit more from architectural modifications (e.g. stronger spatial encoding) rather than fine-tuning technique. These findings suggest that while our Oracle-RLAIF framework excels in improving temporal and action-based comprehension, it is not as effective in improving abstract or spatial tasks. However overall, the general performance gains strongly validate our framework's efficacy in multi-modal video understanding via rank-based feedback.248

6 Related Works

Our work builds on a growing body of literature exploring alignment through preferences, rankings, and human feedback. Most prior approaches to learning from preference signals rely on scalar reward modeling or supervised objectives, limiting their ability to use key rank intuitions during optimization. The core innovation of Oracle-RLAIF lies in directly leveraging rank signal and hierarchical ordering in our proposed loss function, without pre-processing or transformation. While prior work has utilized rating-based feedback [Luu et al., 2025, Xu and Zhu, 2025, Yuan et al., 2023, Choi et al., 2024], none formulate a loss function as we do. Existing approaches differ fundamentally in how they treat ranking information and structure their optimization objectives. For instance, ERL-VLM [Luu et al., 2025] uses rankings to *generate training data for a reward model* rather than to optimize the policy directly. Meta-RL [Xu and Zhu, 2025] maintains ranking across tasks within a meta-learning framework, instead of using group-wise ranks for per-update feedback. RRHF and RAFT [Yuan et al., 2023, Yao et al., 2024] apply ranking in supervised fine-tuning to align with human preferences but do not incorporate it into policy gradient methods. Similarly, LiPO [Zheng et al., 2025] uses listwise preference optimization, but its formulation remains within supervised learning and lacks reinforcement learning updates. In contrast, our method introduces the first policy optimization algorithm that directly integrates ranking feedback, allowing $GRPO_{rank}$ to explicitly model ordinal feedback and better exploit rank-based intuitions.

7 Conclusion

In this work, we introduce a novel reinforcement learning from AI rankings framework to fine-tune large video multi-modal models. We use a drop-in Oracle ranking model in place of the trained and calibrated reward model used in past RLAIF frameworks. To handle ranked feedback, we develop a rank adapted Group Relative Policy Optimization algorithm $GRPO_{rank}$. We train and empirically evaluate Oracle-RLAIF across multiple video-text understanding benchmarks and directly outperform leading fine-tuning methods. Future work includes considering additional types of “Oracle" models and evaluating them (such as closed general-purpose commercial models), and evaluating $GRPO_{rank}$ across a wider range of multi-modal tasks such as audio. We will also explore the use of multi-modal oracles in the context of accelerating general RL tasks [Silva et al., 2017, 2020].

Acknowledgment

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 25-SI-00. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. LLNL-CONF-2010677.

References

Daechul Ahn, Yujin Hu, Oleksandr Ostapenko, Malika Prabhudesai, Cindy Wang, Katerina Fragkiadaki, and Svetlana Lazebnik. Tuning large multimodal models for videos using reinforcement learning from ai feedback. In *ACL*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Alex Fang Chang, Berk Ustun, Hal Ehrenberg, Chuang Gan, and Ser-Nam Lim. Data determines distributional robustness in contrastive language image pre-training (clip). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 648–661. Association for Computational Linguistics, 2021.

Yanjun Chen, Dawei Zhu, Yirong Sun, Xinghao Chen, Wei Zhang, and Xiaoyu Shen. The accuracy paradox in rlhf: When better reward models don’t yield better language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2980–2989, Miami, Florida, USA, 2024. Association for Computational Linguistics.

Heewoong Choi, Sangwon Jung, Hongjoon Ahn, and Taesup Moon. Listwise reward estimation for offline preference-based reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 8651–8671, 2024.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115, 2023.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jiwoo Hong, Noah Lee, Eunki Kim, Guijin Son, Woojin Chung, Aman Gupta, Shao Tang, and James Thorne. On the robustness of reward models for language model alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19894–19918. PMLR, 2023a.

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*, 2023b.

Shuailin Li, Yuang Zhang, Yucheng Zhao, Qiuyue Wang, Fan Jia, Yingfei Liu, and Tiancai Wang. Vlm-eval: A general evaluation on video large language models. *arXiv preprint arXiv:2311.11865*, 2023c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ruyang Liu, Chen Li, Yixiao Ge, Ying Shan, Thomas H. Li, and Ge Li. One for all: Video conversation is feasible without video instruction tuning. In *CVPR (Poster)*, 2024. arXiv preprint arXiv:2309.15785.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

Tung M Luu, Younghwan Lee, Donghood Lee, Sunho Kim, Min Jun Kim, and Chang D. Yoo. Enhancing rating-based reinforcement learning to effectively leverage feedback from large vision-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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.

OpenAI. Gpt-4 technical report. `https://openai.com/research/gpt-4`, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Improving reinforcement learning from human feedback using contrastive rewards. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Felipe Leno da Silva and Anna Helena Reali Costa. A survey on transfer learning for multiagent reinforcement learning systems. *Journal of Artificial Intelligence Research*, 64:645–703, 2019.

Felipe Leno Da Silva, Ruben Glatt, and Anna Helena Reali Costa. Simultaneously learning and advising in multiagent reinforcement learning. In *Proceedings of the 16th conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1100–1108, 2017.

Felipe Leno Da Silva, Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. Uncertainty-aware action advising for deep reinforcement learning agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5792–5799, 2020.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

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

Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning. In *Frontiers of Multimedia Research*, pages 3–29. ACM, 2017.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296, 2016.

Siyuan Xu and Minghui Zhu. Meta-reinforcement learning with adaptation from human feedback via preference-order-preserving task embedding. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Chujie Yao, Runzhe Deng, Yujia Wu, Yichi Zhang, Hu Xu, Yuwei Zhang, Xiang Zhan, Meisam Shah, Ashish Kapoor, Weinan Shi, et al. Raft: Reward-ranked fine-tuning for generative foundation model alignment. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. Oral Presentation.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. In *Advances in Neural Information Processing Systems*, volume 36, pages 10935–10950, 2023.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR (Poster)*, 2024. arXiv preprint arXiv:2303.16199.

Haiyang Zheng, Yuxuan Shao, Baolin Wang, Xiang Ren, Chen Sun, Junnan Liu, Xinyang Liu, and Percy Liang. Listwise preference optimization for language model fine-tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand, 2025. Association for Computational Linguistics.

A Appendix

A.1 VLM-RLAIF Specific Training Configurations

We build directly on the RLAIF framework by [Ahn et al., 2024] which has proven successful in fine tuning *video–language* models. Below we discuss their specific framework framework.

Policy architecture Their policy starts from LLaMA-2-7B and inserts a frozen openai–clip-vit-large-patch14-336 followed by a 32-token Q-FORMER adapter, yielding ≈ 7.1 B parameters. The vision encoder contains two linear layers and additional learnable parameters using LoRA [Hu et al., 2022] This creates a LLaVA framework as the pretrained model [Liu et al., 2023].

Supervised warm-start Before RL, the model is SFT-pretrained on synthetically generated video-text instruction-tuning data (80k) [Maaz et al., 2023, Su et al., 2023], video question answering datasets (67k) [Xiao et al., 2021, Liu et al., 2023], and further generated object-centric datasets (180k). In SFT, the tuning dataset is split into easy (214k) and hard (113k) to create a curriculum learning strategy [Chang et al., 2021] which begins with basic video question answering before progressing to advanced comprehension question answering; where difficulty is measured as the correct answer length.

Reward Model Trained with Context In order to ground their reward model used to score responses for PPO, the authors trained using video, question, video narrative, and preference. This video narrative was gathered by prompting VLM-SFT over the ActivityNet dataset (more than 99,000 videos). The final 13B RM is trained to score responses, and used to drive reinforcement learning with PPO.

Reinforcement Learning with PPO fine-tuning In the VLM-RLAIF fine tuning pipeline, the initial policy outputs two responses per query and computes advantage $\hat{A}_t$ with a learned value head $V_\phi(s_t)$, and updates the policy using the clipped PPO objective and a frozen reference policy (Eq 1).

Training Details For input, videos are uniformly sampled to 50 frames, where the CLIP visual encoder extracts spatial and temporal features similar to [Maaz et al., 2023]. In SFT, LoRA rank and α are set to 32 and training is for one epoch at each stage. For RL, QLoRA [Dettmers et al., 2023] rank is 65 and α to 16 and train the policy model for one epoch . All using $8\text{xNVIDIA A100 GPUS (80G)}$

A.2 Illustrative Examples

Below is a hypothetical table showing computed values for different predicted ranks when $K = 5$, ground-truth rank = 0:

Table 3: $\text{GRPO}_{\text{rank}}$ advantage calculation with ground truth rank = 0 and predicted ranks from 0 to 4

Pred. Rank	**DCG**	**nDCG**	$\boldsymbol{\delta_i = 1 -}$ **nDCG**$_i$	$\mathbb{E}[\boldsymbol{\delta_j}]$	$\hat{\boldsymbol{A}}_{\text{rank}}$
0	1.0000	1.0000	0.0000	0.2887	$+0.2887$
1	0.7925	0.7925	0.2075	0.2887	$+0.0812$
2	0.6667	0.6667	0.3333	0.2887	-0.0446
3	0.5805	0.5805	0.4195	0.2887	-0.1308
4	0.5170	0.5170	0.4830	0.2887	-0.1943