Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback

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

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

{dcahn,yoorachoi,yjy}@yonsei.ac.kr dongyeop@umn.edu jonghyunchoi@snu.ac.kr

Abstract

Recent advancements in large language models have influenced the development of video large multimodal models (VLMMs). Previous approaches for VLMMs involve Supervised Fine-Tuning (SFT) with instruction-tuned datasets, integrating LLM with visual encoders, and additional learnable parameters. Here, aligning video with text, and vice versa, remains a challenge, primarily due to the insufficient quality and quantity of multimodal instruction-tune data compared to that of text-only. This discrepancy often results in alignments that poorly ground the video content. To address this, we present a novel alignment strategy that employs a multimodal AI system equipped with Reinforcement Learning from AI Feedback (RLAIF), providing self-preference feedback to refine itself and facilitating the alignment of video and text modalities. Our approach uniquely integrates detailed video descriptions as context into a multimodal AI system during preference feedback generation to enrich the understanding of video content, a process we call *context-aware reward modeling*. Empirical evaluations on various video benchmarks demonstrate that our VLM-RLAIF outperforms existing approaches, including the SFT model. We commit to open-sourcing our code, models, and datasets to foster further research in this area. https://github.com/yonseivnl/vlm-rlaif

1 Introduction

Large language models (LLMs) are advancing many language and multimodal AI tasks, including those involved with video large multimodal models (VLMMs) (Li et al., 2023b; Muhammad Maaz and Khan, 2023; Lin et al., 2023). Extending the logical reasoning and advanced cognitive capabilities of LLMs to the visual domain, VLMMs are now remarkably proficient in tasks such as video understanding (Li et al., 2023b), video question answering (Ko et al., 2023) and instruction-following

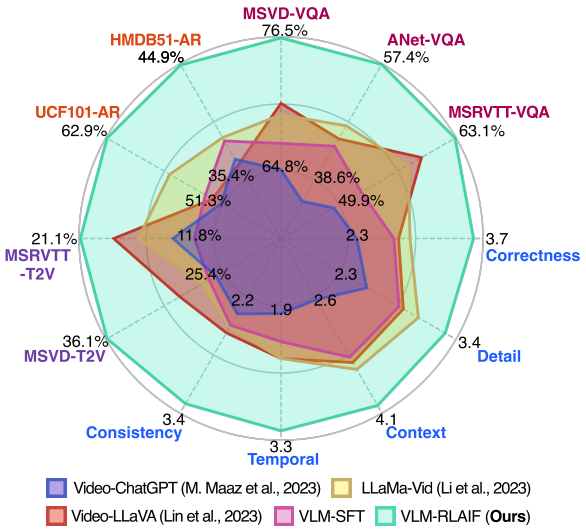

Figure 1: **Quantitative comparison of VLMMs on various video benchmarks.** The video question answering (VQA) task is marked in purple, video-based generative task in blue, the text-to-video (T2V) retrieval task in violet and the action recognition (AR) task in orange color. VLM-RLAIF achieves superior performances on a broad range of video benchmarks compared to previous approaches, including VLM-SFT. Comprehensive comparisons are provided in Tables 1, 2 and 3.

tasks (Muhammad Maaz and Khan, 2023; Luo et al., 2023). These models include publicly available LLMs (Touvron et al., 2023; Chiang et al., 2023; Taori et al., 2023) with visual encoders and additional learnable parameters (Hu et al., 2022; Liu et al., 2023b; Li et al., 2023a). To adapt LLMs to the video modality, thus improving their ability to interpret visual content, they all undergo a supervised fine-tuning (SFT) stage using multimodal instruction-tune data (Luo et al., 2023; Muhammad Maaz and Khan, 2023; Li et al., 2023b).

However, multimodal alignment between video and text faces a significant challenge of deficiency in volume and quality of multimodal instruction-tune data compared to text-only data; text-only data are typically abundant and diverse, while multimodal data are often limited in both quantity and

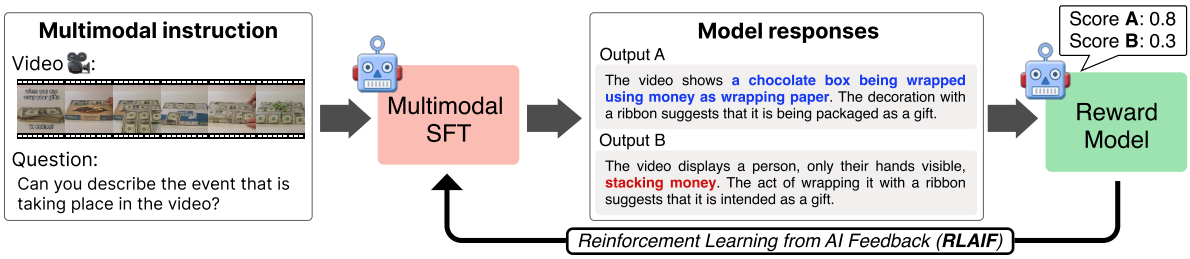

Figure 2: **Illustration of the proposed VLM-RLAIF.** An LLM tuned with video-text instruction-following data, *i.e.*, multimodal SFT model, often produces responses that are not temporally and visually grounded to the video input, as depicted in red color. We propose a method that involves using the VLMM to supervise itself by providing self-preference feedback of generated responses using reward model, refining itself and facilitating the alignment of video and text modalities.

comprehensiveness (Wei et al., 2021; Liu et al., 2023a). This often leads the VLMMs to generate responses that are not properly grounded in the visual content, as illustrated by the SFT model responses in Fig. 2.

To address the issue of VLMMs producing inadequately grounded response, we propose a novel method to align video with text that involves using the VLMM to supervise itself by providing preference feedback of generated responses, as shown in Fig. 2. Specifically, we propose to use Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022; Lee et al., 2023) for multimodal alignment. Unlike the Reinforcement Learning with Human Feedback (RLHF), which has been successful in aligning text-only or image-text based AI agents with human intentions (Ouyang et al., 2022; Sun et al., 2023a; Yu et al., 2023), the RLAIF allows for *scalable oversight* with minimal human intervention. In particular, we improve AI feedback by leveraging detailed video descriptions as a context during the generation of preference feedback, focusing on improved clarity in video content, a process we refer to as *context-aware reward modeling* (Sec. 3.1). In addition, to compensate for the limited multimodal instruction-tune data for training the SFT, we augment it with a human-labeled video question answering and an object-centric multimodal instruction-tune dataset. Further, to effectively utilize this expanded dataset, we propose a simple curriculum training strategy for enhancing the alignment between video and text modality (Sec. 3.2).

We call our proposed framework of training an VLMM with AI feedback as Video Large Multimodal model with RLAIF or **VLM-RLAIF** for short. Our empirical studies show that our aligned VLM-RLAIF exhibits superior performance compared to state-of-the-art VLMMs across a wide array of video benchmarks, as illustrated in Fig. 1.

We summarize our contributions as follows:

- We propose a novel alignment method for video large multimodal models (VLMMs), utilizing Reinforcement Learning from AI feedback (RLAIF) to align video-text modalities effectively.
- We enhance AI's feedback by proposing context-aware reward modeling, focusing on improved clarity and comprehension in video.
- We enrich the SFT model's training by introducing additional instruction-tune data and applying a simple curriculum strategy.
- We demonstrate the effectiveness of our proposed VLM-RLAIF on various video understanding benchmarks by a noticeable margin.

2 Related Work

Multimodal large model. Recently, there have been significant advances for LLMs to go beyond natural language understanding, extending into the realm of multimodal comprehension. The goal is to develop LLMs capable of understanding various modalities, *e.g.*, image (Liu et al., 2023a), video (Li et al., 2023b; Lin et al., 2023), 3D point-cloud (Guo et al., 2023) and *etc*.

To make the LLMs multimodal, most of the work utilize a pretrained encoder, such as CLIP (Radford et al., 2021a), Q-former (Li et al., 2022) or ImageBind (Girdhar et al., 2023), to extract each modality's representations from data. These representations are then projected into the token embedding space of the language model. Then, the models undergo supervised fine-tuning (SFT) with synthetically generated, modality-specific instruction-following datasets. These approaches, adopted

in LLaVA (Liu et al., 2023a), Video-LLaVA (Lin et al., 2023) or Point-LLM (Guo et al., 2023), facilitate the development of proficient conversations grounded in additional modality.

Reinforcement learning from feedback. To operate the model safely and in accordance with human intentions, Reinforcement Learning from Human Feedback (RLHF) has been proposed as a viable solution (Ouyang et al., 2022; Sun et al., 2023a). By collecting preferences from human evaluators, it usually trains the reward model that gives a high reward to the preferred output of the model. However, a significant challenge in this process is the annotation cost associated with selecting the preference. To mitigate this issue, Reinforcement Learning from AI Feedback (RLAIF) was proposed (Bai et al., 2022; Lee et al., 2023; Sun et al., 2023b). RLAIF capitalizes on the inherent ability of Large Language Models (LLMs) to evaluate the generated responses from the SFT model, allowing the LLM itself to assign preferences.

3 VLM-RLAIF Framework

To overcome the limited scalability of human feedback in RLHF, we use AI's feedback to align multimodality between video and text, reducing the reliance on exhaustive human-annotated preferences (Ouyang et al., 2022; Sun et al., 2023a). In particular, we improve the feedback process by using detailed video descriptions, thereby achieving better contextual clarity in video content. The training procedure of VLM-RLAIF can be summarized into three stages as follows:

Supervised fine-tuning (SFT). We first fine-tune an LLM, *e.g.*, Vicuna, using supervised learning on synthetically generated video-text instruction-tune data (Muhammad Maaz and Khan, 2023). This involves the integration of a vision encoder with two linear layers and additional learnable parameters using LoRA (Hu et al., 2022), into the training process. This fine-tuning allows the model to better follow the instructions (Muhammad Maaz and Khan, 2023; Su et al., 2023). Additionally, we improve the SFT process by expanding the instruction-tune data and introducing simple curriculum learning (Sec. 3.2). We refer to this fine-tuned model as the Video Large Multimodal model with SFT or **VLM-SFT** for short.

Reward modeling with AI feedback. A key aspect of the RLAIF involves leveraging a pre-trained AI model to generate human-like preferences between different responses generated from the same input (Bai et al., 2022; Sun et al., 2023b; Lee et al., 2023). To obtain human-like preference, we employ the VLM-SFT as a judge to assess preferences. Once preferences are judged, we train a reward model (RM) based on preferences using a cross-entropy loss, following the Bradley-Terry model for estimating score functions from pairwise preferences (Ouyang et al., 2022; Sun et al., 2023a). We describe the training procedure for collecting preferences and training the reward model in Sec. 3.1. The RM give higher score reward to the better response and lower score reward to the less appropriate one in a pair of responses (see examples in Appendix Fig. 12), thus guiding the policy model using reinforcement learning (RL).

Reinforcement learning from AI feedback. We finally fine-tune a supervised policy model, initialized from the VLM-SFT, aiming to optimize the scalar reward output of the trained RM by reinforcement learning. Specifically, we use the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017), following (Ouyang et al., 2022; Sun et al., 2023b,a).

3.1 Context-Aware Reward Modeling

For VLM-SFT to select preference grounded on the video, we argue that a detailed understanding of video content is necessary for more accurate and contextually relevant decisions by the VLM-SFT. However, the current video encoder presents challenges in accurately encoding the temporal details of videos as they are based on the image encoder (Radford et al., 2021b).

Context-aware preference selection. We propose to explicitly integrate detailed video descriptions, referred to as *context*, into the preference selection workflow, thereby imparting additional contextual clarity to the VLMM, as illustrated in Figures 3-(2) and 4. Specifically, we start by segmenting the video into small clips, each containing up to 20 frames, and then employ the VLM-SFT to generate a detailed video description for each segment with an input prompt, 'Describe this video in detail'. Subsequently, these individual captions are concatenated, which we call a *narrative* of the video. The narrative is then provided to a judge model, *i.e.*, VLM-SFT, for better preference selection. The context not only improves the VLM-SFT's ability

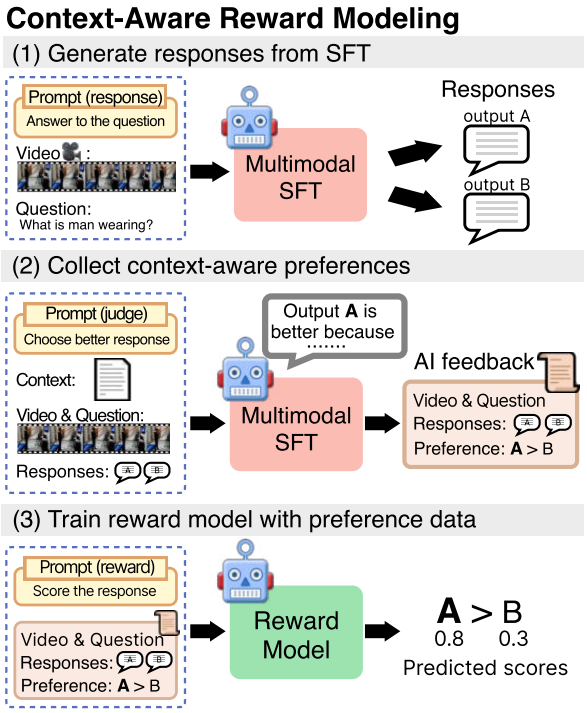

Figure 3: **The three stages of the proposed context-aware reward modeling.** The work flow of each stage is as follows: (1) The SFT model produces two candidate responses from the provided video and question. (2) With the video, question and responses at hand, the SFT model utilize context information and guiding prompt to evaluate the responses. (3) The RM is trained using the preference pairs generated in the previous step as indicated in orange box. Dotted box in each stage denotes a model's input: the first is for generating responses using the SFT model, the second is for the judge model to evaluate and choose the superior response between options A and B, and the third is for training the RM. Each input includes a task-specific prompt, denoted by the yellow box, tailored to guide the model's operation within its respective function (more in Appx. Sec. B).

to obtain a comprehensive view of the video content, but also enables it to identify the most suitable response for the video (see Sec. 4.3 for empirical results). Integrating the context with instruction inputs with a specific prompt (rules for generating preferences as illustrated in Appendix Fig. 9), marked in dotted boxes in Fig. 3-(2), allows us to collect context-aware preferences.

Training the reward model. We design the reward model (RM) to assign higher scores to responses considered *better* and lower scores to those considered *worse* in quality. Starting from VLM-SFT with the final linear layer removed, we train a model to take an input prompt and response, *i.e.*, marked in dotted boxes of Fig. 3-(3), and output a scalar reward value. Using the preference dataset produced from the VLM-SFT, we train the RM with a cross-entropy loss. Specifically, we use 13B VLM-SFT to train the reward model, as it gives slightly better performance than 7B VLM-SFT (see the quantitative comparison discussed in Sec. 4.3). Note that, after training RLAIF using the RM, the 7B VLM-RLAIF significantly surpass the 13B VLM-SFT, thus validating the effectiveness of our proposed framework in aligning video and text modalities (refer to Sec. 4.2 for more details).

3.2 Two-stage Curriculum SFT

During the SFT process, we initially train the LLM with a open-sourced video-text instruction-tune dataset (Muhammad Maaz and Khan, 2023). To improve the VLMM, we not only augment our training with additional video-text instruction-tune datasets but also propose a novel curriculum learning strategy.

Augmenting video instruction-tune dataset. To improve the video understanding ability, we first augment the video-text instruction-tune dataset (Muhammad Maaz and Khan, 2023) with existing human-annotated video question answering datasets (Xiao et al., 2021; Li et al., 2020). In particular, we focus on obtaining instruction-tune dataset that encompass both visual and temporal intricacies for video comprehension. To obtaining such dataset, we use Video Localized Narratives (Voigtlaender et al., 2023), a dataset that provides detailed descriptions of the appearance and complex actions of all objects in the video. We transform the narrative dataset into an object-centric instruction-tune dataset format using ChatGPT similar to prior work (Muhammad Maaz and Khan, 2023). More details about the generated instruction-tune dataset are in Appendix Sec. A.

Curriculum-based supervised fine-tuning. Inspired by the human learning process, we argue the importance of the training sequence, the curriculum learning principles (Chang et al., 2021) for learning VLMMs. We propose a simple two-stage curriculum learning strategy for SFT, aiming to optimize the instruction-following proficiency of VLMMs.

For the curriculum based structured learning, we divide the SFT dataset based on complexity into two segments: *'easy'* and *'hard'*. The model is first trained on *'easy'* tasks to learn basic concepts, then progressed to *'hard'* tasks for advanced challenges.

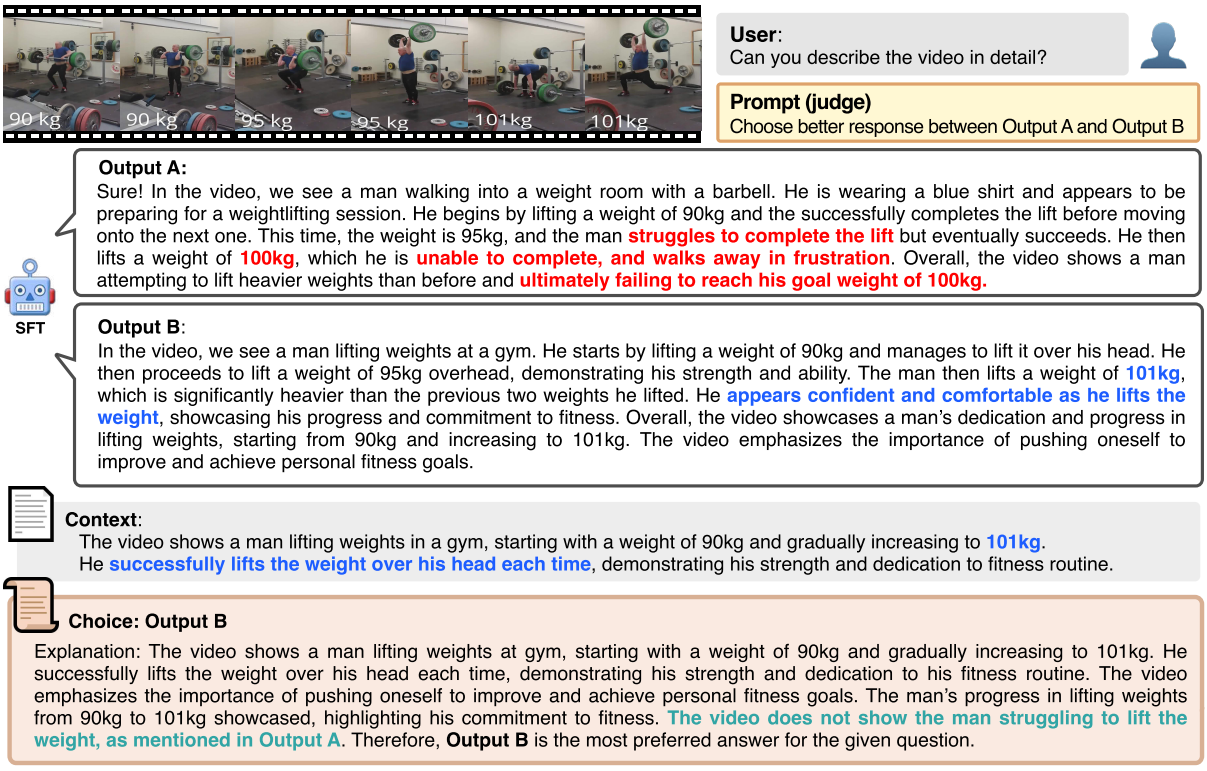

Figure 4: **An example of context-aware preference selection.** We demonstrate our model's ability to generate preference feedback, *i.e.*, preferred choice and explanation marked in orange box, on given instruction input, prompt, two responses and context. Red color denotes an incorrect response, while blue color signifies a correctly grounded response with respect to the visual input. The rationale behind selecting 'Output B' as the preferred choice is indicated in green.

For the difficulty or easiness of the data, given that longer answers often require deeper comprehension of the context and enhanced proficiency in creating syntactically complex sentences (Xu et al., 2020; Agrawal and Singh, 2023; Ranaldi et al., 2023), we use *answer length* as our criterion for sample difficulty, *i.e.*, the longer the answer sentence, the more difficult the task is.

4 Experiments

4.1 Experimental Setup

Model details. We initiate training of the VLM-SFT, building on a pre-trained image-text model (Liu et al., 2023a), with various video-text instruction-tune datasets. In particular, we employ a video projection layer consisting of two linear layers with ReLU activation function between them. Upon establishing the VLM-SFT, we train the RM using the VLM-SFT for both its foundation and the generation of preference data. Subsequently, we train the RM using reinforcement learning (VLM-RLAIF). The policy model is initialized from the VLM-SFT, while the value model is initialized from the RM.

Dataset details. For the SFT dataset, we utilize the open-source video-text instruction-tune dataset (80k) (Muhammad Maaz and Khan, 2023; Li et al., 2023b) and video question answering datasets (67k) (Xiao et al., 2021; Li et al., 2020). More importantly, we generate object-centric narrative video-text instruction-tune dataset (180k) for training the VLM-SFT (Sec. 3.2). For the two-stage curriculum learning, we divide the instruction-tune data for SFT into two groups based on the difficulty; easy (214k) and hard (113k) data. To train the RM, we first generate responses from existing instruction-tune data (Muhammad Maaz and Khan, 2023) and generate preferences using them (40k). Then, we again use the existing instruction-tune dataset (100k) (Muhammad Maaz and Khan, 2023) for RL with the trained RM.

Training details. For the video input, we uniformly sample 50 frames from each video and extract spatial and temporal features from them using CLIP visual encoder, similar to (Muhammad Maaz and Khan, 2023). In the two-stage SFT, we set both the LoRA rank and α to 32, respectively, and train the VLM-SFT for one epoch at each stage.

Methods	LLM Size	Video-based Generative Performance				
		Correctness ↑	Detail ↑	Context ↑	Temporal ↑	Consistency ↑
VideoChat (Li et al., 2023b)	7B	2.23	2.50	2.53	1.94	2.24
LLaMA-Adapter (Zhang et al., 2023b)	7B	2.03	2.32	2.30	1.98	2.15
VideoLLaMA (Zhang et al., 2023a)	7B	1.96	2.18	2.16	1.82	1.79
Video-ChatGPT (Muhammad Maaz and Khan, 2023)	7B	2.40	2.52	2.62	1.98	2.37
Valley (Luo et al., 2023)	7B	2.43	2.13	2.86	2.04	2.45
BT-Adapter (Liu et al., 2023b)	7B	2.68	2.69	3.27	2.34	2.46
VTimeLLM (Huang et al., 2023)	7B	2.78	3.10	3.40	2.49	2.47
Video-LLaVA† (Lin et al., 2023)	7B	2.84	2.86	3.44	2.46	2.57
VideoChat2 (Li et al., 2024)	7B	3.02	2.88	3.51	2.66	2.81
LLaMA-VID (Li et al., 2023d)	7B	2.96	3.00	3.53	2.46	2.51
LLaMA-VID (Li et al., 2023d)	13B	3.07	3.05	3.60	2.58	2.63
GPT-4V (OpenAI et al., 2023)	-	3.85	3.45	3.84	3.63	2.8
VLM-SFT	7B	2.79	2.82	3.37	2.28	2.49
VLM-RLAIF	7B	**3.63**	**3.25**	**4.00**	**3.23**	**3.32**
Δ (RLAIF - SFT)	-	***+0.84***	***+0.43***	***+0.63***	***+0.95***	***+0.83***

Table 1: **Quantitative comparison between different VLMMs on video-based generative performance benchmark.** Our approach, VLM-RLAIF, shows a performance improvement over previous approaches, with the exception of GPT-4V which requires much more computational resource than ours, and demonstrates noticeable enhancements across five criteria when compared to the VLM-SFT. Here, Δ (RLAIF - SFT) indicates the improvement of RLAIF model over SFT model. † denotes reproduced results using the author's implementation.

For RL, we use QLoRA (Dettmers et al., 2023), following (Sun et al., 2023a), setting the rank to 64 and α 16 for computational efficiency and train the policy model for one epoch. All models are trained using 8×NVIDIA A100 GPUs (80G).

4.2 Quantitative Analysis

We evaluate our proposed VLM-RLAIF on various video benchmarks including video-based generative benchmark, zero-shot video question answering (Muhammad Maaz and Khan, 2023; Lin et al., 2023; Li et al., 2023d), text-to-video retrieval, and action recognition (Li et al., 2023c).

Video-based generative performance. We evaluate VLMMs on the video-based generative performance benchmark (Muhammad Maaz and Khan, 2023) that measures five criteria of generated text. In specific, these assess the relevance of the model's output to the video content, its capacity to capture essential details and contextual information, its understanding of temporal sequences, and the consistency in responding to varied yet related queries. As shown in Tab. 1, the VLM-RLAIF performs *on par with* GPT-4V (OpenAI et al., 2023), which requires much more computational resources than ours (*i.e.*, not a fair comparison), and outperforms previous approaches and the VLM-SFT.

Zero-shot video question answering. To evaluate the reasoning ability of VLLMs, we conduct a quantitative evaluation of video question answering (VideoQA) abilities on three datasets (Xu et al., 2017; Yu et al., 2019), following (Muhammad Maaz and Khan, 2023).

The results, as shown in Table 2, indicate that the VLM-RLAIF significantly outperforms previous approaches, including VLM-SFT. Notably, VLM-RLAIF exceeds VLM-SFT by 9.2%, 10.6%, and 13.2% in accuracy and by 0.4, 0.4, and 0.3 in score across all datasets. We believe that the better visually-aligned response generated from the VLM-RLAIF improves the performance (see quantitative analysis in Figures 6 and 11).

Zero-shot text-to-video retrieval. For this task, we follow the procedure proposed in (Li et al., 2023c), which compares CLIP score between generated description and ground-truth caption. Table 3 illustrates the summarized performance comparison to various VLMMs. In the two datasets, *i.e.*, MSVD and MSRVTT, the proposed VLM-RLAIF clearly outperforms other methods including our VLM-SFT by the help of better alignment by the proposed components.

Zero-shot action recognition. Following the VLMMs evaluation procedure proposed in (Li et al., 2023c), we conduct zero-shot action recognition task using two benchmark datasets, *e.g.*, UCF101 and HMDB51. We summarize results of various VLMMs in Tab. 3. In two datasets, the proposed VLM-RLAIF again clearly outperforms other methods including our VLM-SFT.

Methods	LLM Size	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
		Acc.	Score	Acc.	Score	Acc.	Score
FrozenBiLM (Yang et al., 2022)	1B	32.2	-	16.8	-	24.7	-
VideoChat (Li et al., 2023b)	7B	56.3	2.8	45.0	2.5	26.5	2.2
LLaMA-Adapter (Zhang et al., 2023b)	7B	54.9	3.1	43.8	2.7	34.2	2.7
VideoLLaMA (Zhang et al., 2023a)	7B	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT (Muhammad Maaz and Khan, 2023)	7B	64.9	3.3	49.3	2.9	35.2	2.7
Valley (Luo et al., 2023)	7B	60.5	3.3	51.1	2.9	45.1	3.2
BT-Adapter (Liu et al., 2023b)	7B	67.5	3.7	57.0	3.2	45.7	3.2
Video-LLaVA (Lin et al., 2023)	7B	70.7	3.9	59.2	**3.5**	45.3	3.3
VideoChat2 (Li et al., 2024)	7B	70.0	3.9	54.1	3.3	49.1	3.3
LLaMA-VID (Li et al., 2023d)	7B	69.7	3.7	57.7	3.2	47.4	3.3
LLaMA-VID (Li et al., 2023d)	13B	70.0	3.7	58.9	3.2	47.5	3.3
VLM-SFT	7B	67.2	3.6	52.4	3.0	44.1	3.2
VLM-RLAIF	7B	**76.4**	**4.0**	**63.0**	3.4	**57.3**	**3.5**
Δ (RLAIF - SFT)	-	***+9.2%***	***+0.4***	***+10.6%***	***+0.4***	***+13.2%***	***+0.3***

Table 2: **Quantitative comparison between different VLMMs on zero-shot video question answering benchmark.** VLM-RLAIF outperforms previous work across three video-question answering benchmarks.

Methods	LLM Size	T2V Retrieval				Action Recognition			
		MSVD		MSRVTT		UCF101		HMDB51	
		R@1	R@5	R@1	R@5	Top-1	Top-5	Top-1	Top-5
Video-ChatGPT† (Muhammad Maaz and Khan, 2023)	7B	26.03	51.25	14.60	33.80	51.49	79.25	37.10	63.97
Video-LLaVA† (Lin et al., 2023)	7B	29.34	55.35	18.70	38.60	52.33	80.86	36.64	64.03
LLaMA-VID† (Li et al., 2023d)	7B	27.28	53.40	17.00	35.10	56.58	82.79	38.85	65.27
VLM-SFT	7B	26.65	54.27	13.10	30.50	53.03	80.34	38.58	62.37
VLM-RLAIF	7B	**36.03**	**63.40**	**21.00**	**40.70**	**62.83**	**85.86**	**44.75**	**68.37**
Δ (RLAIF - SFT)	-	***+9.38***	***+9.13***	***+7.90***	***+10.2***	***+9.80***	***+5.52***	***+8.11***	***+6.00***

Table 3: **Quantitative comparison between different VLMMs on zero-shot text-to-video (T2V) retrieval and action recognition.** Following (Li et al., 2023c), we evaluate our proposed VLM-RLAIF on zero-shot T2V retrieval and action recognition. †: reproduced by the authors' implementation.

SFT datasets			Curr. learning	Video-based Generative Performance				
[A]	[B]	[C]		Corr. ↑	Det. ↑	Cont. ↑	Temp. ↑	Cons. ↑
✓	✗	✗	✗	2.32	2.53	3.03	2.16	2.23
✓	✓	✓	✗	2.43	2.56	3.09	2.19	2.19
✓	✓	✓	✓	**2.79**	**2.82**	**3.37**	**2.28**	**2.49**

Table 4: **In-depth analysis for the VLM-SFT training procedure.** '[A]' indicate the multimodal instruction-tune dataset proposed in (Muhammad Maaz and Khan, 2023; Li et al., 2023b). '[B]' represents the use of a human-labeled video question answering dataset (Xiao et al., 2021; Li et al., 2020), while '[C]' refers to the use of an object-centric video narrative instruction-tune dataset (Appendix Sec. A). 'Curr. learning' indicates the curriculum learning (Sec. 3.2).

RLAIF	Context Info.	#Clips	Video-based Generative Performance				
			Corr. ↑	Det. ↑	Cont. ↑	Temp. ↑	Cons. ↑
✗	✗	-	2.79	2.82	3.37	2.28	2.49
✓	✗	-	3.26	3.11	3.74	2.78	3.14
✓	✓	1	3.44	3.20	3.89	2.97	**3.36**
✓	✓	3	**3.63**	**3.25**	**4.00**	**3.23**	3.32

Table 5: **Effect of context information on video-based generative performance benchmark.** We investigate the efficacy of using context information for reward modeling (Sec. 3.1). 'Context Info.' indicates the use of context in preference selection. '# Clips' denotes the number of segments into which we divide the video to generate the context information.

4.3 Detailed Analysis

For a detailed analysis, we use the video-based generative benchmark (Muhammad Maaz and Khan, 2023) specifically, as it is well suited to evaluate the wide-ranging capabilities of VLMM, *i.e.*, focusing on response relevance, detail and context capture, temporal understanding, and consistency across queries.

In-depth analysis of SFT training. We first empirically support the effectiveness of augmenting the SFT dataset with additional instruction-following dataset (Sec. 3.2). The first and second rows of Tab. 4 illustrate the benefits of incorporating this additional dataset in improving performance. On top of that, the application of curriculum learning significantly improves performance, implying the efficacy of curriculum learning for the SFT process (the third row of Tab. 4).

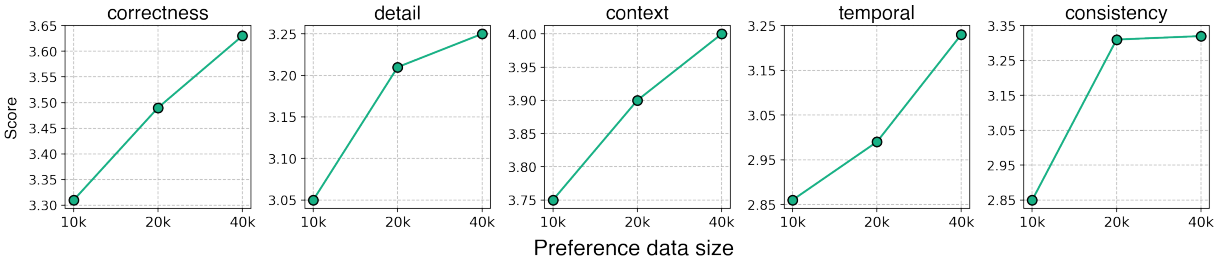

Figure 5: **Effect of preference data size on video-based generative benchmark.** VLM-RLAIF's performance improves across five metrics as the amount of collected preference data increases. The metrics evaluate whether the model generates correct descriptions of the video, includes detailed explanations, remains contextual, demonstrates temporal understanding, and provides consistent responses to variations of the same question.

Methods	LLM Size	Video-based Generative Perf.				
		Corr. ↑	Det. ↑	Cont. ↑	Temp. ↑	Cons. ↑
VLM-SFT	7B	2.32	2.53	3.03	2.16	2.23
VLM-SFT	13B	**2.64**	**2.73**	**3.28**	**2.38**	**2.44**

Table 6: **Quantitative comparison between different sizes of the VLM-SFT.** We assess the performance of VLM-SFT with varying LMM sizes, specifically 7B and 13B, on video-based generative benchmarks. We conduct this evaluation without the integration of augmented instruction-tune data and the implementation of two-stage curriculum learning.

Effect of preference data size. Our method's strength lies in generating synthetic preference feedback in large quantities. To study the benefit of large-sized data, we sweep the size of preference data from 10k – the same quantity utilized in the multimodal image-text RLHF framework (Sun et al., 2023a) – to 40k. As expected, we observe monotonic increases in performance by the increase in data size, as shown in Figure 5.

Context-aware preference labeling. Table 5 presents a study on the impact of context information. This reveals that the use of context information during preference selection improves performance, as demonstrated in the second and third rows in Tab. 5. Moreover, segmenting the video into multiple clips, specifically three, and using their aggregated captions as context further improves performance (see the third and fourth rows). We conjecture that segmenting video into clips with detailed captions enables the model to capture detailed nuances, thereby enriching the contextual clarity for reward modeling.

Comparison between different SFT model sizes. Table 6 shows a clear performance margin between VLM-SFT with different LLM sizes, specifically 7B and 13B, in video-based generative tasks. Larger LLM sizes lead to improved performance,

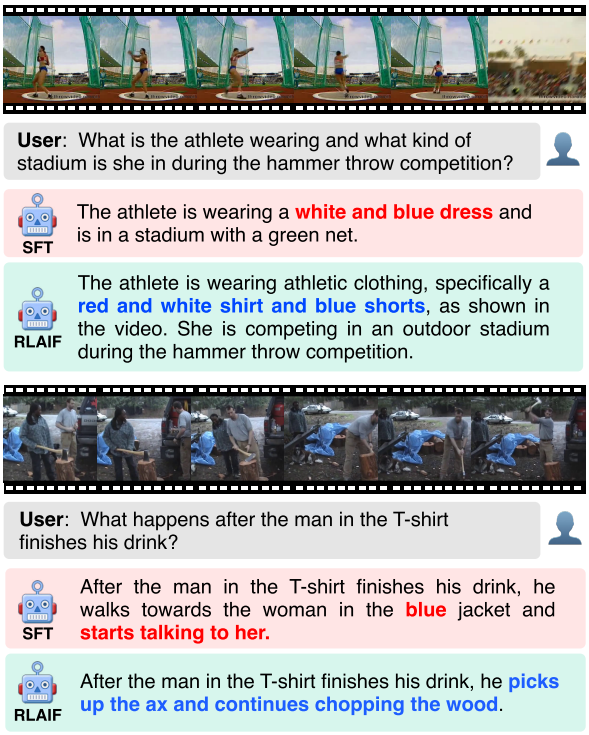

Figure 6: **Samples of VLM-RLAIF and VLM-SFT in video understanding.** Two samples demonstrate better capability of the VLM-RLAIF. Red color indicates visually un-grounded expressions considering video content and blue color indicates well grounded expressions to the video.

suggesting that increased model capacity enhances the ability to capture and generate complex video content. Thus, we adopt the 13B model for the reward model (Sec. 3.1), with supporting evidence in Tables 7 and 8.

Various LLM size for reward model Tables 7 and 8 show the performance of VLM-RLAIF with different RM sizes, initialized from VLM-RLAIF-7B and VLM-RLAIF-13B. In both cases, the policy model is initialized with VLM-SFT-7B. Our RLAIF method outperforms VLM-SFT significantly across all evaluation methods. Specifically,

Methods	LLM Size		Video Question Answering						Video-based Generative Perf.				
			MSVD		MSRVTT		ActivityNet						
	Policy Model	Reward Model	Acc.	Score	Acc.	Score	Acc.	Score	Corr. ↑	Det. ↑	Cont. ↑	Temp. ↑	Cons. ↑
LLaMA-VID (7B)	-	-	69.7	3.7	57.7	3.2	47.4	3.3	2.96	3.00	3.53	2.46	2.51
VLM-SFT	7B	-	67.2	3.6	52.4	3.0	44.1	3.2	2.79	2.82	3.37	2.28	2.49
VLM-RLAIF	7B	7B	75.1	3.9	61.0	3.3	56.1	3.4	3.47	3.14	3.87	3.05	3.30
VLM-RLAIF	7B	13B	**76.4**	**4.0**	**63.0**	**3.4**	**57.3**	**3.5**	**3.63**	**3.25**	**4.00**	**3.23**	**3.32**

Table 7: **Quantitative comparison between different size of policy model and reward model for the VLM-RLAIF.** We evaluate the VLM-RLAIF with different model size for policy model and reward model on zero-shot video question answering and video-based generative benchmark.

Methods	LLM Size		T2V Retieval				Action Recognition			
			MSVD		MSRVTT		UCF101		HMDB51	
	Policy Model	Reward Model	R@1	R@5	R@1	R@5	Top-1	Top-5	Top-1	Top-5
LLaMA-VID (7B)	-	-	27.28	53.40	17.00	35.10	56.58	82.79	38.85	65.27
VLM-SFT	7B	-	26.65	54.27	13.10	30.50	53.03	80.34	38.58	62.37
VLM-RLAIF	7B	7B	33.73	61.95	20.80	**42.90**	61.09	85.15	43.86	65.88
VLM-RLAIF	7B	13B	**36.03**	**63.40**	**21.00**	40.70	**62.83**	**85.86**	**44.75**	**68.37**

Table 8: **Quantitative comparison between different size of policy model and reward model for the VLM-RLAIF.** We evaluate the VLM-RLAIF with different model size for policy model and reward model on zero-shot text-to-video retrieval and action recognition tasks.

RLAIF with 7B RM achieves a 5-12% improvement in zero-shot video question answering. Scaling up the RM from 7B to 13B further improves performance, except for the text-to-video retrieval task R@5 metric on MSR-VTT.

4.4 Qualitative Analysis

We now qualitatively compares the performance of VLM-SFT and VLM-RLAIF, highlighting their multimodal understanding capabilities in Figure 6. VLM-RLAIF consistently yields more accurate answers than VLM-SFT, as shown in the detailed recognition of the attire of an athlete in the first example, marked in red and blue. The second example further affirms VLM-RLAIF's benefit in generating better grounded responses to the visual input, where VLM-SFT falls short. More examples are in the Appendix Fig. 11 for the space sake.

5 Conclusion

We propose a novel alignment strategy for VLMMs, termed VLM-RLAIF, that uses reinforcement learning from AI feedback. To improve multimodal alignment, we propose a context-aware reward modeling, enabling AI to generate feedback for self-improvement. In addition, we expand the instruction-tune dataset for SFT and adopt a curriculum-based training approach, that is particularly effective in the gradual learning of complex video-text relationships. In our empirical validations, the VLM-RLAIF significantly outperforms previous models in multiple multimodal video-text understanding benchmarks, which implies good generalization performance across tasks.

Limitations

Given that our approach utilizes feedback synthesized by the AI model, the effectiveness of our proposed VLM-RLAIF largely depends on the quality of the AI model's generated responses. In light of recent studies exploring the use of artificially generated data (Koo et al., 2023; Das et al., 2024), we believe that there needs further research to enhance the quality of synthetically generated data, thereby establishing a more reliable RLAIF system.

In addition, although we have evaluate our model across a range of benchmarks for VLMMs, *e.g.*, videoQA, video-based generative tasks, retrieval, and recognition tasks, there are other tasks such as temporal reasoning (Liang et al., 2022) that are necessary for VLMMs to be effectively applied in real-world scenarios. Application of our method to these tasks would be a great future research avenue.