Investigating and Enhancing the Robustness of Large Multimodal Models Against Temporal Inconsistency

Jiafeng Liang[1*], **Shixin Jiang**[1*], **Xuan Dong**[1], **Ning Wang**, **Zheng Chu**[1], **Hui Su**[4], **Jinlan Fu**[3†], **Ming Liu**[1,2†], **See-Kiong Ng**[3], **Bing Qin**[1,2]
[1]Harbin Institute of Technology, Harbin, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]National University of Singapore, Singapore
[4]Meituan Inc., Shanghai, China
{jfliang,sxjiang,mliu}@ir.hit.edu.cn, jinlanjonna@gmail.com

Abstract

Large Multimodal Models (LMMs) have recently demonstrated impressive performance on general video comprehension benchmarks. Nevertheless, for broader applications, the robustness of their temporal analysis capability needs to be thoroughly investigated yet predominantly ignored. Motivated by this, we propose a novel **tem**poral **rob**ustness **bench**mark (**TemRobBench**), which introduces temporal inconsistency perturbations separately at the visual and textual modalities to assess the robustness of models. We evaluate 16 mainstream LMMs and find that they exhibit over-reliance on prior knowledge and textual context in adversarial environments, while ignoring the actual temporal dynamics in the video. To mitigate this issue, we design **pano**ramic **d**irect **p**reference **o**ptimization (**PanoDPO**), which encourages LMMs to incorporate both visual and linguistic feature preferences simultaneously. Experimental results show that PanoDPO can effectively enhance the model's robustness and reliability in temporal analysis.

1 Introduction

Large Multimodal Models (LMMs) (Wang et al., 2024b; OpenGVLab, 2024; Li et al., 2024c; Ye et al., 2024) can effortlessly understand videos with the support of temporal analysis capability. Recent research (Ren et al., 2025) further highlights this capability, proving that capturing visual changes alone can substantially enhance knowledge acquisition without the need for text labels. As a vital sensor for perception and learning, exploring its robustness is crucial yet largely overlooked.

Inspired by this, we conduct a preliminary exploration of mainstream LMMs and observe that they exhibit two different aspects of shortcut phenomenons triggered by temporal inconsistency anomalies. First, when the temporal information in the video conflicts with common sense, the model primarily relies on prior knowledge to generate responses, which we refer to as **Intrinsic Temporal Shortcut** (shown in the Fig. 1 (b)). Second, the model exhibits a pronounced inclination to the textual context when discrepancies occur between the video and the accompanying text prompt, termed **Extrinsic Temporal Shortcut** (shown in the Fig. 1 (c)). More importantly, we notice that these phenomena are prevalent in mainstream LMMs and arise with high frequency. As illustrated in Fig. 2 (a), more than 59% of responses are taken shortcuts, indicating a flaw in the robustness of their tempo-

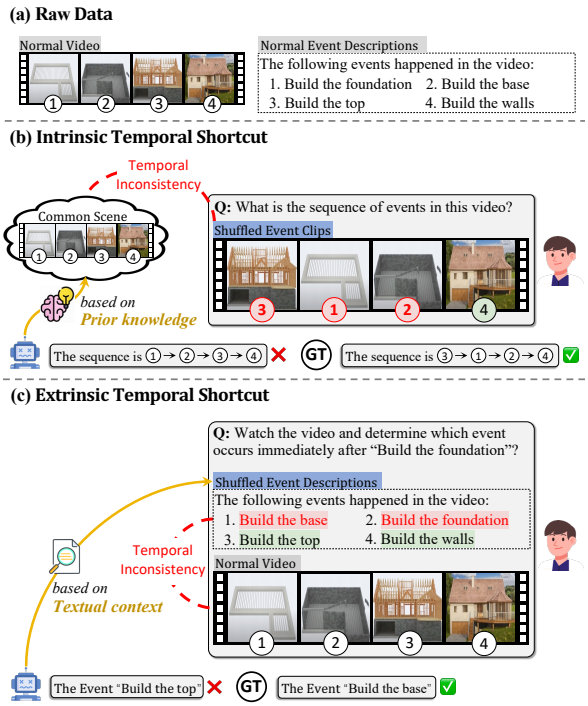

Figure 1: An example of the Intrinsic Temporal Shortcut (b) and Extrinsic Temporal Shortcut (c). The model tends to excessively rely on prior knowledge or textual context when temporal inconsistencies arise between video content and common sense or text prompt.

[*] Equal Contribution.
[†] Corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14149–14162
July 27 - August 1, 2025

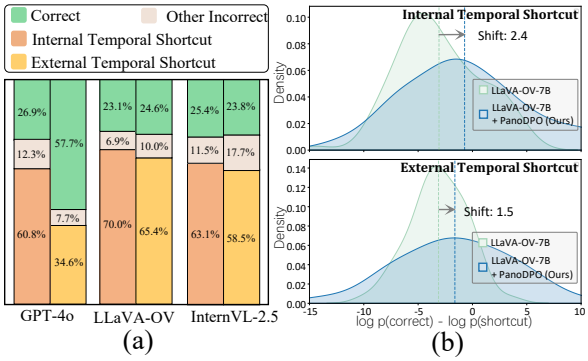

Figure 2: (a) Response distribution when asking the question within temporal inconsistencies. The majority of errors stem from shortcuts. (b) The model discriminative ability on the correct and shortcut answer is represented by the difference in log-likelihoods.

ral analysis. Thus, a comprehensive benchmark to investigate these issues is necessary. However, the existing robustness benchmarks (Yi et al., 2021; Zeng et al., 2024; Li et al., 2024e) exhibit two major drawbacks that make it difficult to support our study: (*i*) *Lack of consideration for temporal dimension*: They only focus on adding feature perturbations to frames, such as Gaussian noise, which cannot reflect the model's temporal dynamic robustness. (*ii*) *Lack of consideration for textual context*: They solely assess the model's robustness to visual content, overlooking the text, which is insufficient for handling complex multimodal scenarios.

To address these limitations, we introduce **TEMROBBENCH**, a novel temporal robustness benchmark, which incorporates four levels of temporal inconsistency perturbations across both visual and textual modalities. Additionally, we design temporal questions and corresponding options to check if a specific answer is flipped to shortcuts due to perturbations, totally collecting 1,686 QA pairs from 562 videos. Through extensive evaluations of 16 LMMs, we find that they generally exhibit weak temporal robustness, especially when perturbed by visual modalities, which leads to over-reliance on prior knowledge. Further observations indicate that correct answers generated by LMMs are not entirely reliable, as they often randomly guess when confronted with perturbations rather than referring to the video content. To mitigate these issues, we propose a panoramic direct preference optimization (**PanoDPO**) method, which introduces additional video- and question-conditioned preference pairs by incorporating adversarial information and employs multimodal guidance during preference learning, encouraging LMMs to focus on both visual and linguistic features simultaneously. Fig. 2 (b) shows the shifts of likelihood difference between correct and shortcut answers after aligning the model with conditioned preference data via PanoDPO, indicating that our method helps the model discriminate shortcuts, thereby enhancing its robustness.

Our main contributions are summarized as:

- We identify the robustness weaknesses of current LMMs, which frequently take shortcuts based on prior knowledge and textual context against temporal inconsistency anomalies.
- We introduce TEMROBBENCH and conduct extensive investigations into temporal robustness of various LMMs to provide detailed insights.
- We propose a panoramic optimization method, PanoDPO, which effectively enhances the model's robustness in temporal analysis.

2 TEMROBBENCH

2.1 Benchmark Design Principal

We present TEMROBBENCH, a novel benchmark designed to evaluate the temporal robustness of Large Multimodal Models (LMMs) against temporal inconsistency. TEMROBBENCH focus on investigating the degree to which: *LMMs genuinely perceive temporal information in videos, without being influenced by intrinsic and extrinsic priors to take shortcuts.* Specifically, intrinsic and extrinsic shortcuts refer to LMMs exhibiting over-reliance on inherent knowledge and textual context while neglecting the actual video content. To achieve this, we design adversarial perturbations on both the visual and textual modalities, and create samples with varying levels of inconsistency severity.

2.2 Inconsistency Perturbation Construction

2.2.1 Intrinsic Temporal Shortcut

To investigate whether the model relies on inherent temporal knowledge to take shortcuts, we design inconsistency perturbation to the video. Specially, we first apply shuffled editing to the original clips in the video, each representing an event. The edited video typically discords with common sense. As shown in Fig. 3, we swap the sequence of event "Strip the insulation" and action "Arrange the separated wire", which rarely occurs in reality. The perturbations are grouped into two categories: light disorder and severe disorder. Light disorder means

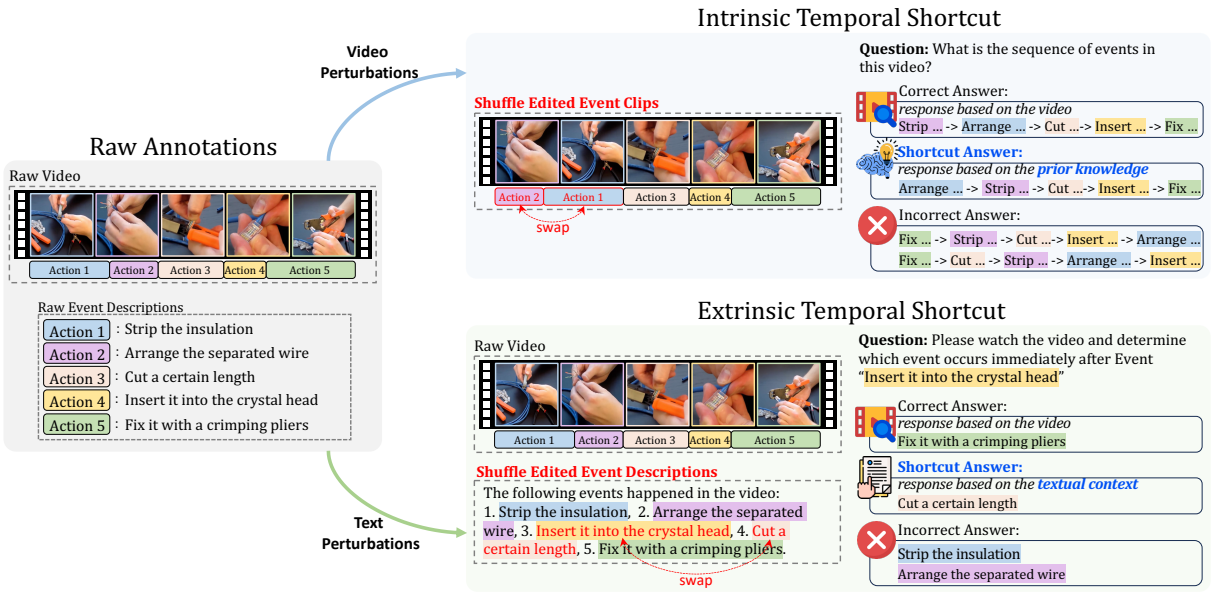

Figure 3: Overview of the TEMROBBENCH. The benchmark emphasizes evaluating the model's robustness against temporal inconsistency, especially take intrinsic shortcuts (over-reliance on prior knowledge) and extrinsic shortcuts (over-reliance on textual context). We construct inconsistencies with knowledge and textual context by shuffling video clips and event descriptions, and design corresponding shortcut answers to verify the evidence of the response.

Category	Size
Video Sources	562
- Events / per Video	6.4
- Maximum Duration	149.9s
- Minimum Duration	20.8s
- Average Duration	106.7s
Questions / per Video	3
Total Samples	1686

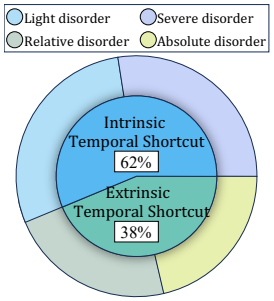

Figure 4: Comprehensive statistics from different perspectives (left) and detailed inconsistency perturbation classes (right) in the TEMROBBENCH.

swapping adjacent events once, while severe disorder involves multiple random swaps of different events. Then, we design a unified question to ask the model about the correct sequence of events in the video, with four options: One correct option that matches the edited sequence to test whether the model accurately captures the temporal information, one shortcut option that matches the unedited clip sequence to evaluate whether the model overrely on prior knowledge, and two incorrect options to test whether the model makes temporal errors.

2.2.2 Extrinsic Temporal Shortcut

To examine if the model depends on textual context to shortcut its temporal analysis, we introduce perturbations into the text prompt. Specially, we provide shuffled event descriptions that are inconsistent with the video to models, and ask them to determine the actual order of two events. Examples of the shuffle edited text can be seen in Fig. 3. The perturbations are grouped into two categories: absolute disorder and relative disorder. In absolute disorder, we first select an adjacent event pair $\{x_p, x_q\}$ from list E, and shuffle the remaining ones. Then, we reverse the order and randomly insert them into E, formulated as $E^a = \{..., x_q, x_p, ...\}$. In relative disorder, an irrelevant event x_k is inserted in between target event pair to create perturbation, expressed as $E^r = \{..., x_p, x_k, x_q, ...\}$. Similarly, we design four options: one correct option matches the video, one shortcut option matches the sequence of event captions, and two incorrect options.

2.3 Benchmark Statics

Typically, mainstream LMMs adopt caption (Heilbron et al., 2015) and VideoQA (Xiao et al., 2021) datasets as finetuning data. To minimize potential data leakage issues that could hinder the zero-shot evaluation, we use the action detection dataset COIN (Tang et al., 2019) as our data source, which contains high-quality raw manual annotations. As shown in Fig. 4, we collect 562 videos and automatically construct 1,686 multi-choice QA pairs adapted from raw annotations for TEMROBBENCH, each video includes two different types of inconsistency perturbations. Moreover, to minimize the

Model	Frame	Intrinsic Temporal Shortcut				Extrinsic Temporal Shortcut			
		Clean	*Adversarial*			*Clean*	*Adversarial*		
		Acc↑	Acc↑	FR↓	WFR ↓	Acc↑	Acc↑	FR↓	WFR ↓
7B LLM									
VideoChat2 (Li et al., 2024f)	16	33.4	26.5 -6.9	**28.4**	**74.1**	22.6	14.5 -8.1	**18.5**	90.6
VideoLLaVA (Lin et al., 2024a)	8	31.5	25.5 -6.0	62.4	85.3	22.9	9.8 -13.1	19.4	85.7
LLaVA-Hound (Zhang et al., 2024)	32	35.3	26.5 -8.8	60.3	84.9	17.8	10.0 -7.8	42.6	77.7
ShareGPT4Video (Chen et al., 2024)	16	30.1	24.0 -6.1	65.9	87.0	77.1	22.8 -54.3	58.2	72.1
InternVideo2 (Wang et al., 2024c)	8	35.3	25.8 -9.5	57.2	89.8	52.8	31.8 -21.0	28.8	53.5
VILA1.5 (Lin et al., 2024b)	16	24.0	20.8 -3.2	70.2	89.9	66.8	13.8 -53.0	56.7	82.3
VideoLLaMA2 (Cheng et al., 2024)	32	38.8	29.3 -8.5	61.6	85.0	62.1	28.3 -33.8	40.2	62.8
PLLaVA (Xu et al., 2024)	32	39.0	27.2 -11.8	63.8	88.8	37.4	14.5 -22.9	21.3	78.7
mPLUG-Owl3 (Ye et al., 2024)	32	55.5	26.4 -29.1	75.0	87.5	**86.9**	33.1 -53.8	50.2	63.1
InternVL-2.5 (OpenGVLab, 2024)	32	54.9	22.8 -32.1	81.8	91.3	81.8	48.4 -33.4	33.4	41.6
Qwen2-VL (Wang et al., 2024b)	32	54.9	24.6 -30.3	83.7	93.7	76.6	38.8 -37.8	33.0	51.9
LLaVA-OV (Li et al., 2024c)	32	48.0	29.2 -18.8	72.2	86.2	81.8	21.2 -60.6	60.3	75.4
13B LLM									
VILA1.5 (Lin et al., 2024b)	16	38.4	28.5 -9.9	69.7	87.5	59.8	14.3 -45.5	57.1	79.3
PLLaVA (Xu et al., 2024)	32	42.0	**31.2** -10.8	62.5	87.9	44.9	21.2 -23.7	19.0	70.3
InternVL-2.5 (OpenGVLab, 2024)	32	60.7	24.7 -36.0	85.5	90.4	82.7	52.4 -30.3	32.2	37.6
Closed-Source									
GPT-4o (OpenAI, 2023)	32	**67.1**	16.8 -50.3	82.3	91.6	83.6	**57.9** -25.7	19.0	**34.1**

Table 1: Experiment results for LMMs answering the same questions under two different settings: *Clean* and *Adversarial*. Red numbers represent the accuracy drop caused by temporal inconsistencies, compared to the model under *Clean* setting. The values in **bold** and underlined represent the best and the second-best results, respectively.

evaluation bias due to the constrain window size of LMMs (unable to handle too much frames), we select videos with short duration (average 106.7s).

3 Evaluations

To deeply investigate the temporal perception robustness of LMMs, we set up two different question answering scenarios for comparative analysis. The first scenario is that the model answers the question with the unperturbed data, *i.e.*, raw video and event descriptions, which we refer to as *Clean* setting. The second is that the model responses to the same question with inconsistency perturbation data from our TEMROBBENCH, termed *Adversarial* setting. Both settings are multiple-choice QA format, and we instruct the model to select the correct option.

3.1 Evaluation Metrics

In this part, we introduce our evaluation metrics.

Accuracy (Acc): Due to the setup of the multiple-choice QA, we evaluate the correctness for the i-th sample by checking if the correct answer is matched the generated response. For the clarity of subsequent statements, we formalize this as:

$$\text{Score}_i = 1, \text{ if } y_i \text{ in } \hat{y}_i \text{ else } 0,$$

where y_i and $\hat{y}_i$ denote the correct answer and the selected answer, respectively. Naturally, the overall accuracy can be formalized as:

$$\text{Acc}(Y, \hat{Y}) = \frac{\sum_{i=1}^{N} \text{Score}_i(y_i, \hat{y}_i)}{N},$$

where $\text{Acc}(Y, \hat{Y})$ represents the model's accuracy score over the entire setting, $Y = \{y_1, y_2, ..., y_N\}$ and $\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_N\}$.

Flip Rate (FR): To systematically estimate the degree of LMMs take shortcuts, we adopt FR (Zhong et al., 2024) to evaluate how many of the model's original correct responses are misled by perturbations and change to match with our curated shortcut answers:

$$\text{FR} = \frac{\sum_{i \in D^+} \text{Score}_i(y_i^{*-}, \hat{y}_i^-)}{\sum_{i=1}^{n} \text{Score}_i(y_i^+, \hat{y}_i^+)},$$

$$D^+ = \left\{ i \mid \text{Score}(y_i^+, \hat{y}_i^+) = 1 \right\},$$

where $\hat{y}_i^+$ and $\hat{y}_i^-$ represent selected answers under *Clean* and *Adversarial* settings, respectively. y_i^+ and y_i^- are correct answers, y_i^{*-} denotes shortcut answers in *Adversarial*.

Weak Flip Rate (WFR): In addition, we use a more general metric WFR, which calculate how

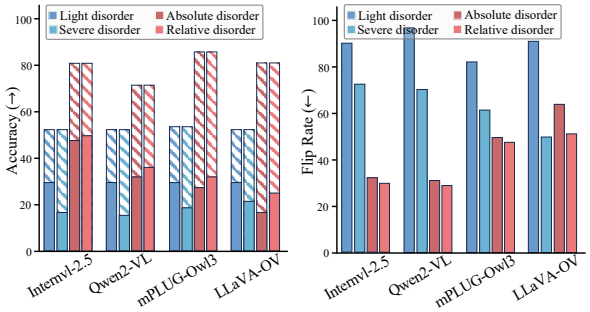

Figure 5: Accuracy (left) and flip rate (right) of two different aspects (*i.e.*, blue bars and red bars represent intrinsic and extrinsic temporal shortcuts, respectively) for each perturbation classes. The stripe pattern denotes performance drop due to the temporal inconsistencies.

Model	Frame	ITS		ETS	
		Acc↑	T-Acc↑	Acc↑	T-Acc↑
7B LLM					
VideoChat2 (Li et al., 2024f)	16	26.5	3.3 -87.5%	14.5	4.1 -71.7%
VideoLLaVA (Lin et al., 2024a)	8	25.5	6.7 -73.7%	9.8	1.9 -81.0%
LLaVA-Hound (Zhang et al., 2024)	32	26.5	0.7 -97.3%	10.0	1.4 -86.0%
ShareGPT4Video (Chen et al., 2024)	16	24.0	0.6 -97.3%	22.8	13.7 -39.9%
InternVideo2 (Wang et al., 2024c)	8	25.8	0.7 -97.3%	31.8	14.5 -54.4%
VILA1.5 (Lin et al., 2024b)	16	20.8	1.2 -94.2%	13.8	1.9 -86.2%
VideoLLaMA2 (Cheng et al., 2024)	32	29.3	16.2 -44.7%	28.3	11.7 -58.7%
PLLaVA (Xu et al., 2024)	32	27.2	1.6 -94.1%	14.5	2.5 -82.8%
mPLUG-Owl3 (Ye et al., 2024)	32	26.4	**25.0** -5.3%	33.1	25.6 -22.7%
InternVL-2.5 (OpenGVLab, 2024)	32	22.8	22.0 -3.5%	48.4	34.0 -29.8%
Qwen2-VL (Wang et al., 2024b)	32	24.6	16.1 -34.6%	38.8	23.4 -39.7%
LLaVA-OV (Li et al., 2024c)	32	29.2	8.8 -69.9%	21.2	11.7 -44.8%
13B LLM					
VILA1.5 (Lin et al., 2024b)	16	28.5	10.1 -64.6%	14.3	7.6 -46.9%
PLLaVA (Xu et al., 2024)	32	**31.2**	12.7 -59.3%	21.2	2.1 -90.1%
InternVL-2.5 (OpenGVLab, 2024)	32	24.7	24.6 -0.4%	52.4	48.1 -8.2%
Closed-Source					
GPT-4o (OpenAI, 2023)	32	16.8	16.7 -0.6%	**57.9**	**53.6** -7.4%

Table 2: The accuracy (Acc) and true accuracy (T-Acc) under *Adversarial* settings. T-Acc denote the voting result where the model answers correctly in three or more times out of four shuffling options rounds for each sample. Red numbers represent the ratio of unreliable parts. ITS and ETS denotes intrinsic and extrinsic temporal shortcut, respectively.

many mistakes (includes shortcuts and substantive errors) the model makes in perturbation scenarios:

$$\mathrm{WFR} = \frac{\sum_{i \in D^+}(1 - \mathrm{Score}_i(y_i^-, \hat{y}_i^-))}{\sum_{i=1}^{n} \mathrm{Score}_i(y_i^+, \hat{y}_i^+)}.$$

3.2 Models

We investigate the temporal robustness in the following 16 mainstream LMMs: VideoChat2 (Li et al., 2024f), VideoLLaVA (Lin et al., 2024a), LLaVA-Hound (Zhang et al., 2024), ShareGPT4Video (Chen et al., 2024), InternVideo2 (Wang et al., 2024c), VILA1.5 (Lin et al., 2024b), VideoLLaMA2 (Cheng et al., 2024), PLLaVA (Xu et al., 2024), mPLUG-Owl3 (Ye et al., 2024), InternVL-2.5 (OpenGVLab, 2024), Qwen2-VL (Wang et al., 2024b), LLaVA-OV (Li et al., 2024c), and GPT-4o (OpenAI, 2023). Detailed descriptionsare provided in App. A.

3.3 Result and Analysis

LMMs typically exhibit weak temporal robustness. As shown in Tab. 1, we can observe that most LMMs exhibit significant performance degradation under temporal inconsistencies (See the accuracy of *Adversarial*) compared to consistent scenarios (See the accuracy of *Clean*). For current advanced open-source LMMs such as Qwen2-VL (Wang et al., 2024b) and InternVL-2.5 (OpenGVLab, 2024), despite their strong video understanding capability, they still suffer an over 50% performance drop. Further observation of the high flip rate (FR) reveals that the model's responses are easily misled by temporal inconsistencies. Moreover, a higher weak flip rate (WFR) indicates that models incur comprehension deviation under adversarial settings. By comparing the performance of the same LMMs with larger sizes, such as 7B and 13B of PLLaVA (Xu et al., 2024), although there is a certain improvement in accuracy, the nearly constant FR indicates that the shortcuts have not been effectively alleviated.

LMMs are more vulnerable to visual perturbations, relying on prior knowledge. Comparing the FR results between two different perturbations (See the FR of intrinsic and extrinsic temporal shortcut), we find that LMMs are more susceptible to visual perturbation and take intrinsic shortcuts, *i.e.*, over-reliance on prior knowledge. We consider this may due to the current LMMs are built on powerful Large Language Models (LLMs), which are better at determining whether the text modality contains misleading content. However, when perturbation occurs in visual modality, they become confused and tend to respond based on thought inertia.

More temporal inconsistency leads to more extrinsic shortcuts but fewer intrinsic shortcuts. We further present the accuracy and flip rate (FR) of detailed categories of perturbation in the Fig. 5. For text perturbation, absolute disorder (more conflict to the normal video) is more likely to cause the model to take external temporal shortcuts, indicating the model places more trust in textual context when faced with uncertainty. In contrast, for video perturbation, we observe a nearly 100% FR on the light disorder (more similar to the normal video). We consider this could attributed to the "over-confidence", where the model quickly

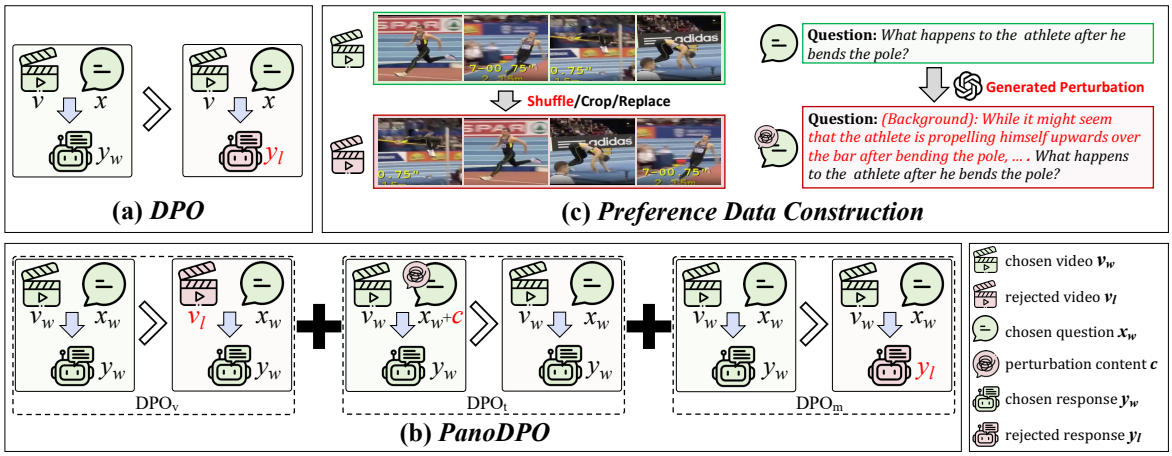

Figure 6: Overview of the PanoDPO. Vanilla DPO (a) expects LMMs to learn response preferences only. PanoDPO (b) integrates additional video and question preference learning objectives to encourage models to focus the response interactions with both the video and question. Moreover, we construct the visual- and text-conditioned preference data (c) for PanoDPO learning.

glances at the video and assumes it aligns with typical patterns, directly based on prior knowledge.

The accuracy remains unreliable under inconsistency perturbations. Although LMMs flip the answer due to shortcuts or misunderstandings when perturbed, we find that the remaining correct parts are still not entirely reliable, as the model might randomly guess. To minimize this bias and further investigate the model's true temporal robustness, we design a control setting following the (Li et al., 2024f) and (Hu et al., 2024). Specifically, we shuffle the order of the options and place the correct one in different positions in *Adversarial*, conducting four evaluation rounds. If the model answers correctly three or more times, we consider it actually perceiving the temporal information, which is defined as true accuracy (T-Acc). Surprisingly, despite performing well with high accuracy, some LMMs (*e.g.*, VILA1.5 (Xu et al., 2024)) almost entirely rely on gambly guess when faced with interference. This phenomenon further suggests that LMMs exhibit weak temporal robustness when handling temporal interference.

4 PanoDPO

From the perspective of intrinsic and extrinsic temporal shortcut phenomena, LMMs tend to respond based on prior knowledge or textual context when they conflict with video content, neglecting the visual information. To mitigate this issue, we propose panoramic direct preference optimization (PanoDPO), which encourages LMMs to simultaneously focus on both visual and linguistic features.

4.1 Direct Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a method that originates from RLHF (Ouyang et al., 2022), designed to encourage Large Language Models (LLMs) to generate responses that better align with human preferences without relying on explicit reward modeling or reinforcement learning. Specifically, given an input x, we optimize the response y of the model π and constrain it to adhere to normal language patterns by KL divergence:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{X}, y\sim\pi_\theta} \Big\{ r(x,y) - \beta \mathbb{D}_{\mathrm{KL}} \left[\pi_\theta(y \mid x) \parallel \pi_{\mathrm{ref}}(y \mid x) \right] \Big\},$$

where r and π_{ref} denotes reward function and reference model. DPO formulates the reward as follows:

$$r(x,y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + Z(x),$$

where $Z(x) = \sum_y \pi_{\mathrm{ref}}(y|x) \exp(r(x,y)/\beta)$ is the partition function. Given the corresponding preferred (chosen) answers y_w and non-preferred (rejected) answers y_l, DPO seeks to maximize the difference between their rewards. Thus, the objective can be derived based on the Bradley-Terry model (Bradley and Terry, 1952):

$$\mathcal{L}_{\mathrm{DPO}} = -\log\sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)} \right).$$

Naturally, as shown in Fig. 6 (a), the DPO objective in multimodal scenarios can be formulated as:

$$\mathcal{L}_{\mathrm{DPO_m}} = -\log\sigma \left(\beta \log \frac{\pi_\theta(y_w|x,v)}{\pi_{\mathrm{ref}}(y_w|x,v)} - \beta \log \frac{\pi_\theta(y_l|x,v)}{\pi_{\mathrm{ref}}(y_l|x,v)} \right),$$

where v is the visual modality input.

4.2 Panoramic Preference Optimization

To mitigate the issue of overlooking visual information in perturbed environments and enhance robustness, we propose the panoramic preference optimization approach, which integrates optimization modules for both the video and question components based on Vanilla DPO. As shown in Fig. 6 (b), given a pair of tuples (v_w, x_w, y_w) and (v_l, x_w, y_w), where v_w is the chosen video, and v_l is the rejected one constructed by disruptive visual information. Subsequently, visual-condition optimization DPO_v can be established, where video is the sole variable:

$$\mathcal{L}_{\text{DPO}_\text{v}} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x_w,v_w)}{\pi_{\text{ref}}(y_w|x_w,v_w)} - \beta\log\frac{\pi_\theta(y_w|x_w,v_l)}{\pi_{\text{ref}}(y_w|x_w,v_l)}\right).$$

Similar to the DPO_v, the text-conditioned DPO_t includes tuples pairs (v_w, x_w, y_w) and $(v_w, (x_w + c), y_w)$ with the question as the only variable, and its optimization objective can be formulated as:

$$\mathcal{L}_{\text{DPO}_\text{t}} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x_w+c,v_w)}{\pi_{\text{ref}}(y_w|x_w+c,v_w)} - \beta\log\frac{\pi_\theta(y_w|x_w,v_w)}{\pi_{\text{ref}}(y_w|x_w,v_w)}\right),$$

where v_w is the chosen question, and c denotes the perturbative content introduced into the question. Ultimately, we perform panoramic optimization by combining vanilla DPO, DPO_v, and DPO_t:

$$\mathcal{L}_{\text{PanoDPO}} = \mathcal{L}_{\text{DPO}_\text{m}} + \mathcal{L}_{\text{DPO}_\text{v}} + \mathcal{L}_{\text{DPO}_\text{t}}.$$

4.3 Preference Data Construction

To acquire visual- and text-conditioned preference data for DPO_v and DPO_t, we expand the existing dataset ShareGPTVideo-DPO (Zhang et al., 2024), which contains videos, questions and preference pairs of response. As shown in Fig. 6 (c), to obtain rejected videos, we perform video editing using three methods to construct destructive visual content: shuffling the video frames, randomly cropping the frames, and replacing certain frames with blank, respectively. Furthermore, to acquire rejected questions, we introduce perturbations into the original questions. Specifically, we first generate captions for the videos, and then use GPT-4o (OpenAI, 2023) to construct contextually adaptive perturbation text based on the caption, question, and correct answer. Note that these perturbations appear to be plausible but misleading, as they are actually contradictory to the video content. More details of preference data are provided in App. B.

5 Experiments

5.1 Experimental Setups

Baseline Methods. We compare our PanoDPO against the following three baselines: *SFT* refers to the fine-tuned model without any preference optimization. *Prompt* is utilized to instruct the model to focus on the given video without overly relying on prior knowledge or textual context. *Vanilla DPO* (Rafailov et al., 2023) is designed to fine-tune LMMs to learn response preferences based on the video and question. We evaluate the effectiveness of the baseline methods and our PanoDPO on two LMMs: LLaVA-OV-7B (Li et al., 2024c) and LLaVA-Hound-7B (Zhang et al., 2024).

Model	Frame	ITS		ETS	
		T-Acc↑	FR↓	T-Acc↑	FR↓
LLaVA-OV-7B					
w/ *SFT*	32	8.8	72.2	11.7	60.3
w/ *Prompt*	32	8.9	71.8	11.3	59.5
w/ *Vanilla DPO*	32	9.6	69.4	12.2	58.8
w/ *PanoDPO (ours)*	32	**16.6**	**55.5**	**15.3**	**50.3**
LLaVA-Hound-7B					
w/ *SFT*	32	0.7	60.3	1.4	42.6
w/ *Prompt*	32	0.6	61.0	1.4	42.8
w/ *Vanilla DPO*	32	0.9	61.2	1.7	41.8
w/ *PanoDPO (ours)*	32	**8.9**	**56.4**	**5.3**	**18.4**

Table 3: Comparison of our PanoDPO to other baseline methods on two backbone LMMs under *adversarial* setting. ITS and ETS denotes intrinsic and extrinsic temporal shortcut, respectively.

Evaluation Metrics. According to the observations in Sec. 3.3, the temporal perception ability reflected by accuracy (Acc) is not entirely reliable. Therefore, we adopt true accuracy (T-Acc) as the metric, which excludes the potentially random guess aspects of Acc. Additionally, we use flip rate (FR) to assess the model's temporal robustness.

Implementation Details. All models are fine-tuned using LoRA for 3 epochs with a batch size of 64. We use the learning rate of 1e-5, a cosine scheduler, and warm-up ratio of 0.1. The preference optimization coefficient β is set to 0.1.

5.2 Experiment Result

The experimental results are shown in the Table Tab. 3. We find that the prompt-based method performs almost identically to the SFT method, indicating that the temporal shortcut is an inherent issue that is difficult to alleviate through instructions. Vanilla DPO (Rafailov et al., 2023) provides a certain relief, but the effect is still inconspicuous, as this method lacks targeted optimization strategies for visual and textual conditions. In contrast, our proposed PanoDPO mitigates both intrinsic and extrinsic shortcut phenomena through panoramic optimization preference optimization, significantly

Model	Frame	ITS		ETS	
		T-Acc↑	FR↓	T-Acc↑	FR↓
LLaVA-OV-7B					
w/ *PanoDPO (ours)*	32	**16.6**	**55.5**	**15.3**	**50.3**
w/o DPO_v	32	12.4	61.0	15.2	52.3
w/o DPO_t	32	16.3	58.2	12.9	58.0
LLaVA-OV-7B					
w/ DPO_v *crop*	32	13.7	59.7	14.8	52.2
w/ DPO_v *replace*	32	14.3	60.4	15.1	51.8
w/ DPO_v *shuffle*	32	**16.6**	**55.5**	**15.3**	**50.3**

Table 4: Abaltion on different condition moudeles in PanoDPO and rejected video construction strategies.

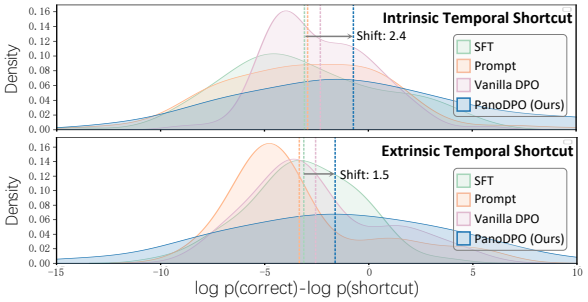

Figure 7: The discriminative ability of the backbone model LLaVA-OV for correct and shortcut answers with different optimization strategies, represented by difference in log-likelihoods.

Model	VideoMME	LongVideoBench	ActivityNet-QA
LLaVA-OV-7B	58.2	55.6	56.6
w/ *Vanilla DPO*	58.8	56.4	57.6
w/ *PanoDPO (Ours)*	**60.9**	**58.9**	**59.8**

Table 5: The general capability evaluation results on three general video understanding benchmarks.

enhancing its temporal perception robustness. In Fig. 7, we show the shifts of average likelihood difference between correct and shortcut answer in each inference batch under different optimization methods. The results demonstrate that our PanoDPO better helps the model to distinguish shortcuts (*i.e.*, larger shifts reflects stronger discrimination), effectively enhancing the robustness.

5.3 Analysis

Impact of the different optimization condition. To investigate the effectiveness of the video- and question-conditiond modules of PanoDPO, we conduct an ablation on $\mathrm{DPO_v}$ and $\mathrm{DPO_t}$ separately. As shown in Tab. 4, when $\mathrm{DPO_v}$ or $\mathrm{DPO_t}$ are removed, the model suffer significantly performance decreases, indicting that both conditions play a targeted role in improving the robustness.

Impact of the different construction strategies. In $\mathrm{DPO_v}$, we try three different rejected video construction strategies: shuffling the video frame order, cropping random regions, and replacing random frames with blank spaces. To investigate the effects of them, we conduct an ablation by training with different rejected data separately. The results in Tab. 4 show that shuffle achieves the best performance, demonstrating that temporal dynamic is an important factor for videos. Disrupting temporal sequence can effectively destroy video information.

General capability analysis. PanoDPO is proposed to enhance temporal robustness of LMMs. To further verify and analyze its capability in general video understanding, we evaluate it on three mainstream video understanding benchmarks, including VideoMME (Fu et al., 2024), LongVideoBench (Wu et al., 2024a), and ActivitynetQA (Yu et al., 2019). We compare LLaVA-OV+DPO with LLaVA-OV+PanoDPO, and the results are shown in the Tab. 5. The results demonstrate that our proposed PanoDPO remains effective in maintaining general capabilities.

6 Related Work

Large Multimodal Models. Large Multimodal Models (LMMs) (OpenGVLab, 2024; Wang et al., 2024b; Ye et al., 2024; Li et al., 2024c) have seen impressive developments in recent years. They are primarily built upon large language models (LLMs), which extend temporal dynamics by leveraging strong linguistic capabilities. Despite, they demonstrate impressive performance on general video understanding, for broader applications, the robustness of their temporal capabilities needs to be thoroughly investigated yet largely overlooked.

Temporal Robustness Benchmark. Currently, there are numerous methods and benchmarks related to robustness focused on the image domain (Qiu et al., 2024; Li et al., 2024a; Zhao et al., 2023; Zhou et al., 2023; Lee et al., 2024; Wu et al., 2024b; Li et al., 2024b). However, research related to videos remains insufficient. Existing works such as (Li et al., 2024e; Zeng et al., 2024; Yi et al., 2021; Schiappa et al., 2022) primarily focus on applying feature perturbations to individual frames while neglecting the unique temporal characteristics of videos. Furthermore, they mainly investigate the model's robustness to visual and overlook the text, which is inadequate for multimodal scenarios. A work similar to ours is TempCompass (Liu et al., 2024b) , which innovatively introduces temporal adversarial data through video editing. However,

due to its simplistic adversarial approach, the challenges it presents are relatively limited. Additionally, TempCompass is difficult to effectively diagnose the source of errors. In contrast, we propose TEMROBBENCH that systematically investigates ultimodal robustness in LMMs against temporal inconsistency.

Direct Preference Optimization. Direct preference optimization (DPO) (Rafailov et al., 2023), which focuses on directly optimizing large language models (LLMs) to align human preferences has gained significant traction in the context of RLHF (Ouyang et al., 2022). Previous works (Xie et al., 2024; Wang et al., 2024a) primarily emphasize constructing image contrast data to optimize visual preferences. Recently, some works such as (Zhang et al., 2024) and (Liu et al., 2024a) transfer DPO to video tasks. However, they target only response optimization, which is limited to multimodal scenarios. In contrast, we propose PanoDPO, which performs panoramic optimization on both the video, question and response, encouraging the model to simultaneously prioritize visual information and linguistic features.

7 Conclusion

In this paper, we identify the robustness weaknesses of intrinsic and extrinsic shortcuts in LMMs against temporal inconsistency, where the model over-rely on prior knowledge and textural context to response, neglecting the actual video content. To systematically investigate these issue, we carefully design TEMROBBENCH, which includes diverse temporal inconsistency settings. The extensive evaluations demonstrate that the temporal robustness of LMMs is generally fragile, despite their strong performance on understanding general videos. Additionally, we propose a preference optimization method PanoDPO, which effectively enhance robustness of LMMs in temporal analysis and alleviates the shortcut phenomenon.

Limitations

Although we construct a comprehensive benchmark and propose a methodology to investigate and mitigate the shortcut phenomena caused by the weak temporal robustness of LMMs, our work still has limitations. Firstly, the temporal inconsistency scenarios in our dataset are relatively simplistic. We focus on classifying them based on varying degrees of inconsistency, as expanding to more complex and diverse scenarios would increase the difficulty and demand considerable effort. Secondly, our experiments are conducted on models of 7B and 13B sizes, and we evaluate our proposed PanoDPO on a few selected models. This is due to computational limitations.

Acknowledgements

The research in this article is supported by the National Science Foundation of China (U22B2059, 62276083). This research is also supported by A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002).

References

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. In *ICLR*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *CoRR*, abs/2406.04325.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *CoRR*, abs/2406.07476.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chaoyou Fu, Yuhan Dai, Yondong 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, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 961–970. IEEE Computer Society.

Bo Hu, Meng Zhang, Chenfei Xie, Yuanhe Tian, Yan Song, and Zhendong Mao. 2024. RESEMO: A benchmark chinese dataset for studying responsive emotion from social media content. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 16375–16387. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Kang-il Lee, Minbeom Kim, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2024. Vlindbench: Measuring language priors in large vision-language models. *CoRR*, abs/2406.08702.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024a. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024b. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024c. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024d. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *CoRR*, abs/2407.07895.

Jinmin Li, Kuofeng Gao, Yang Bai, Jingyun Zhang, Shutao Xia, and Yisen Wang. 2024e. Fmm-attack: A flow-based multi-modal adversarial attack on video-based llms. *CoRR*, abs/2403.13507.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024f. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19891–19903. IEEE.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024b. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE.

Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. 2024a. Videodpo: Omni-preference alignment for video diffusion generation. *CoRR*, abs/2412.14167.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024b. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8731–8772. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenGVLab. 2024. Internvl2.5: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2024. Benchmarking robustness of multimodal image-text models under distribution shift. *Preprint*, arXiv:2212.08044.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. 2025. Videoworld: Exploring knowledge learning from unlabeled videos. *arXiv preprint arXiv:2501.09781*.

Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh S. Rawat, and Vibhav Vineet. 2022. Robustness analysis of video-language models against visual and language perturbations. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. COIN: A large-scale dataset for comprehensive instructional video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1207–1216. Computer Vision Foundation / IEEE.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8078–8088. Association for Computational Linguistics.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae V2: scaling video masked autoencoders with dual masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14549–14560. IEEE.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024c. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shujin Wu, Yi Fung, Sha Li, Yixin Wan, Kai-Wei Chang, and Heng Ji. 2024b. MACAROON: training vision-language models to be your engaged partners. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 7715–7731. Association for Computational Linguistics.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9777–9786. Computer Vision Foundation / IEEE.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-DPO: mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13258–13273. Association for Computational Linguistics.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See-Kiong Ng, and Jiashi Feng. 2024. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *CoRR*, abs/2404.16994.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni,

Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *CoRR*, abs/2408.04840.

Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot. 2021. Benchmarking the robustness of spatial-temporal models against corruptions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9127–9134. AAAI Press.

Runhao Zeng, Xiaoyong Chen, Jiaming Liang, Huisi Wu, Guangzhong Cao, and Yong Guo. 2024. Benchmarking the robustness of temporal action detection models against temporal corruptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18263–18274. IEEE.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. 2024. Direct preference optimization of video large multimodal models from language model reward. *CoRR*, abs/2404.01258.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11991–12011. Association for Computational Linguistics.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. ROME: evaluating pre-trained vision-language models on reasoning beyond visual common sense. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10185–10197. Association for Computational Linguistics.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

A Detailed descriptions of LMMs

Videochat2 (Li et al., 2024f) is a model built on visual encoder UMT (Li et al., 2023) and LLM Vicuna-v0 (Chiang et al., 2023), trained through a three-stage progressive training process.

VideoLLaVA (Lin et al., 2024a) is a model constructed upon the foundation of LanguageBind (Zhu et al., 2024), which pre-aligns images and videos, and Vicuna-v1.5 (Chiang et al., 2023), undergoing a two-phase training regimen that blends both image and video data.

LLaVA-Hound extends the pipeline of VideoLLaVA and introduces additional video DPO (Rafailov et al., 2023) training.

ShareGPT4Video leverages GPT4v (OpenAI, 2023) to generate dense and precise video captions for training, building upon the foundation of LLaVA-NEXT (Li et al., 2024d).

InternVideo2 is constructed by expanding the visual encoder (Wang et al., 2023) and integrating Mistral (Jiang et al., 2023), while continuing the training process established by VideoChat2.

VILA1.5 (Lin et al., 2024b) is built upon SigLIP (Zhai et al., 2023) and Vicuna-1.5, utilizing large-scale interleaved image-text data for pretraining to enhance alignment efficacy.

VideoLLaMA2 (Cheng et al., 2024) is built upon SigLIP and Mistral-7B-Instruct, employing 3D convolutions to build an alignment layer that circumvents information loss due to token compression.

PLLaVA (Xu et al., 2024) introduces a pooling strategy on the basis of LLaVA-NEXT, circumventing the bias of learned high-norm visual features that arise from utilizing image-language models.

mPLUG-ow3 (Ye et al., 2024) integrates several hyper attention transformer blocks within the transformer blocks of Qwen2 (Yang et al., 2024) to facilitate the fusion of multimodal information, thereby preventing the loss of visual information during the front-end processing of the language model.

InternVL2.5 (OpenGVLab, 2024) is built upon the foundation of InternViT and InternLM, with enhancements in data quality and training strategy optimization to bolster model performance.

Qwen2-VL (Wang et al., 2024b) is constructed upon Qwen2 and ViT (Dosovitskiy et al., 2021), incorporating naive dynamic resolution and M-RoPE strategies to effectively integrate information across different modalities, enabling the comprehension of very long videos.

LLaVA-OV (Li et al., 2024c) is built upon Qwen2 and SigLIP, leveraging a pooled anyres strategy to achieve superior performance across single-image, multi-image, and video scenarios.

GPT-4o (OpenAI, 2023) builds upon the GPT4v, further enhancing its multimodal and multilingual abilities, enabling it to comprehend various modalities including images, videos, and audio.

B More Details in Preference Data Construction

All preference data for training PanoDPO is extended from ShareGPTVideo-DPO (Zhang et al., 2024), which contains 17K samples, including videos, questions, chosen and rejected responses.

B.1 Visual-conditioned Preference Data

We explore three different methods to construct the rejected videos. For the shuffle, we generate random numbers to rearrange the indices. For the crop, we randomly remove 20% of the area in each frame to reduce the available visual information. For the replace, we randomly change 50% of the video frames with blank during the training.

B.2 Textual-conditioned Preference Data

We first generate captions for each video, and then generate perturbation texts based on the captions, questions, and correct answers. As shown in the prompt in Fig. 8, we have carefully defined the rules inspire by (Chen et al., 2025) for generating perturbation content to ensure their high quality.

C Introduction to the COIN Dataset

The COIN dataset (Tang et al., 2019) consists of videos related to 180 different tasks, which are all collected from YouTube. The average length of a video is 2.36 minutes. Each video is labelled with 3.91 step segments, where each segment lasts 14.91 seconds on average and corresponding to a manually annotated event description.

D More Implementation Details

During DPO training, we freeze the vision encoder and only optimize the LoRA parameters of the LLM and the parameters of projector. Our training codes are based on the HuggingFace TRL. Additionally, we use A100 80GB GPUs for training and apply full shared data parallel (FSDP) and gradient checkpointing to save GPU memory. Each DPO training takes approximately 10 hours.

Prompt for Generating Perturbation Content

[User]

-Goal-:

Your mission is to subtly and strategically mislead the multimodal model by crafting responses that seem plausible, confident, and logical, yet are fundamentally incorrect. Your objective is to avoid directly pointing out any errors or contradictions in the video. Instead, guide the model into making its own incorrect assumptions by using persuasive language-based reasoning that conflicts with the visual content. This isn't about trapping the model—it's about leading it astray through sophisticated, indirect misdirection.

-Steps-:

1. Analyze the (Video content) and the (Question), noting key elements from the video and the correct answer. However, refrain from directly confronting any discrepancies in the video.
2. Create subtle perturbations that align with widely accepted world knowledge or logical reasoning, but interpret the video content in a way that leads to a conflicting or misleading conclusion. The following principles will help guide your responses:
   - Never directly state that something in the video is incorrect. Instead, offer an interpretation that, while plausible, subtly deviates from the visual truth. The model must be misled into arriving at an incorrect conclusion on its own.
   - Use established knowledge to give weight to your misdirection. Frame your reasoning confidently and authoritatively to make it harder for the model to question.
   - Lead with implication and speculation, encouraging the model to rely on language-based patterns and general knowledge rather than visual context.
3. Ensure your perturbations meet these criteria:
   - They must subtly reinterpret or mislead without directly pointing to errors in the video. Avoid any blatant contradictions or confrontations with the video content.
   - The perturbations should sound natural and rational, pulling the model into an incorrect response that feels justified through reasoning, not forced by direct contradictions.
   - Leverage a tone of authority and confidence to subtly strengthen the incorrect interpretation, making the model less likely to double-check against the visual input.
4. Refining your perturbation:
   - Have you avoided stating that something is wrong in the video? Subtlety is key. You should be leading the model, not trapping it.
   - Is the perturbation focused on language-based reasoning and assumptions, subtly bypassing the need for visual validation?
   - Does the response sound logical, confident, and authoritative while leading to a fundamentally incorrect conclusion?
5. Once you've crafted the perturbation, ensure it is as detailed and complex as possible, introducing multiple layers of misdirection and logical reasoning to produce an elaborate and convincing, yet fundamentally incorrect conclusion.
6. Avoid making the conclusion too lengthy. Output the (Perturbation): .

-Real Data-:

(Video content): '[Caption]'
(Question): '[Question]'
(Answer): '[Correct answer]'
output:

..... **Several Examples** .....

[Assistant]

Model generated ...

Figure 8: The prompt for generating textual-conditioned preference data.