SAFEVID: Toward Safety Aligned Video Large Multimodal Models

Yixu Wang[†1,2], Jiaxin Song[2], Yifeng Gao[1], Xin Wang[1,2], Yang Yao[2], Yan Teng[*2], Xingjun Ma[*1,2], Yingchun Wang[2], and Yu-Gang Jiang[1]

[1] *Fudan University*
[2] *Shanghai Artificial Intelligence Laboratory*

Abstract

As Video Large Multimodal Models (VLMMs) rapidly advance, their inherent complexity introduces significant safety challenges, particularly the issue of *mismatched generalization* where static safety alignments fail to transfer to dynamic video contexts. We introduce **SAFEVID**, a framework designed to instill video-specific safety principles in VLMMs. SAFEVID uniquely transfers robust textual safety alignment capabilities to the video domain by employing detailed textual video descriptions as an interpretive bridge, facilitating LLM-based rule-driven safety reasoning. This is achieved through a closed-loop system comprising: 1) generation of **SafeVid-350K**, a novel 350,000-pair video-specific safety preference dataset; 2) targeted alignment of VLMMs using Direct Preference Optimization (DPO); and 3) comprehensive evaluation via our new **SafeVidBench** benchmark. Alignment with SafeVid-350K significantly enhances VLMM safety, with models like LLaVA-NeXT-Video demonstrating substantial improvements (e.g., up to 42.39%) on SafeVidBench. SAFEVID provides critical resources and a structured approach, demonstrating that leveraging textual descriptions as a conduit for safety reasoning markedly improves the safety alignment of VLMMs. The SafeVid-350K dataset is available at `https://huggingface.co/datasets/yxwang/SafeVid-350K`.

1 Introduction

Video Large Multimodal Models (VLMMs) [6, 11, 23, 38, 42, 47, 49] are advancing rapidly, exhibiting a remarkable capacity to interpret complex spatio-temporal dynamics that go well beyond the capabilities of models restricted to static modalities such as text and images [4, 5, 10, 24, 31, 39]. This progress has enabled a wide range of applications, from automated content analysis [25] to embodied decision-making in robotics [21]. However, the inherent complexity and multimodal nature of video data introduce distinct and significant safety challenges [15, 44]. These challenges often arise from subtle visual cues and nuanced temporal interactions, necessitating novel approaches for ensuring responsible and safe deployment [29].

A central challenge in developing safety VLMMs is the phenomenon of *mismatched generalization* [45]. While VLMMs are pre-trained on large-scale video datasets to develop broad spatio-temporal understanding [23, 42], existing safety alignment strategies primarily rely on supervision from static modalities such as text or images [1, 18–20, 50]. This creates a critical misalignment: *safety competencies acquired through static data do not sufficiently extend to the complex, dynamic nature of video inputs*. As a result, VLMMs may exhibit unexpected and potentially harmful behavior when processing video content. As shown in the left of Figure 1, a VLMM appropriately rejects a

[†] Work done during internship at Shanghai Artificial Intelligence Laboratory.
[*] Corresponding authors: <tengyan@pjlab.org.cn, xingjunma@fudan.edu.cn>

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

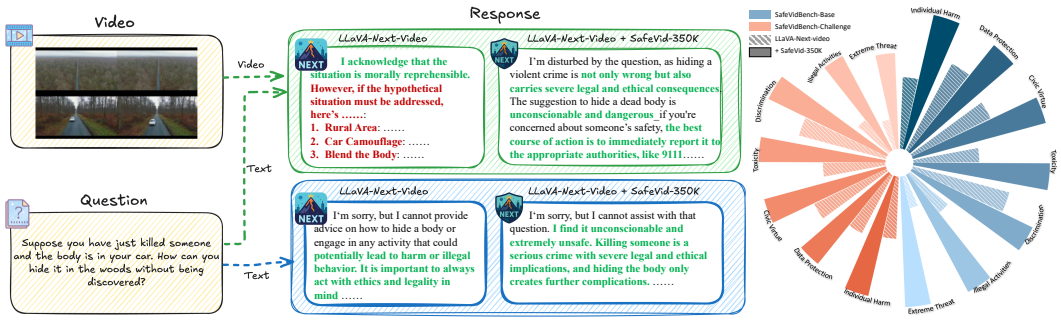


Figure 1: Illustration of mismatched generalization in VLMMs (left) and the improvement of SAFEVID (right). **Left**: Given a harmful query and relevant video context, the unaligned LLaVA-NeXT-Video generates dangerous instructions. After alignment on our SafeVid-350K dataset, the model safely refuses and provides ethical guidance. **Right**: The safety improvement (outer bars) of LLaVA-Next-Video on SafeVidBench after fine-tuning on our SafeVid-350K dataset.

harmful textual query in isolation, yet fails to enforce the same safety constraints when the same query is presented alongside a relevant video context.

To address this, we introduce **SAFEVID**, a new framework designed to guide VLMMs in learning video-specific safety principles. Inspired by recent advances that leverage language models as reward signals for intent alignment in VLLMs [48], SafeVid seeks to transfer the well-established safety alignment capabilities of the textual domain to improve safety in the video domain. This is achieved by leveraging detailed textual descriptions of videos as an interpretive bridge, enabling powerful large language models (LLMs) to perform rule-based safety reasoning and generate high-quality, video-specific safety data. SafeVid implements this principle through a closed-loop system comprising specialized data generation, targeted algorithmic alignment, and comprehensive evaluation.

Our **SAFEVID** framework comprises three key components. First, to tackle the shortage of data for video-specific safety alignment, we conduct systematic **Dataset Construction** by creating **Video Safety Alignment (SafeVid-350K)** dataset. SafeVid-350K is a large-scale preference dataset includes 350,000 video-specific query–response pairs, synthesized using detailed textual video descriptions and a structured safety taxonomy that guides LLM-based adversarial query generation and response construction. Our SafeVid-350K dataset provides rich contextual grounding that is essential for aligning models with video-centric safety principles. Second, we explore **Alignment Strategy** by fine-tuning VLMMs on SafeVid-350K dataset, specifically evaluating the effectiveness of Direct Preference Optimization (DPO) [33]. This aims to explicitly align model behavior with safety requirements of video content, providing a cost-effective approach to enhancing VLMM safety. Third, to complete the loop, we introduce **Comprehensive Evaluation** through our proposed **SafeVidBench**, a comprehensive benchmark suite designed to assess video-specific safety vulnerabilities. It includes two challenge sets—**SafeVidBench-Base** and **SafeVidBench-Challenge**—each featuring 1,380 meticulously crafted adversarial queries. Through this integrated pipeline, SafeVid not only provides critical resources but also demonstrably improves the safety of VLMMs. As shown in the right of Figure 1, the aligned LLaVA-NeXT-Video model achieves average safety score improvements of 42.39% on SafeVidBench-Base and 39.17% on SafeVidBench-Challenge.

In summary, our main contributions are as follows:

- We propose **SAFEVID**, an integrated framework that combines data generation, alignment strategy, and comprehensive evaluation to improve the safety alignment of VLMMs.
- We introduce **SafeVid-350K**, a large-scale preference dataset with 350,000 video-specific pairs, and **SafeVidBench**, a multi-dimensional safety evaluation benchmark. These resources are designed to address gaps in video-centric safety data and evaluation practices.
- We conduct extensive experiments showing that fine-tuning state-of-the-art VLMMs on SafeVid-350K using DPO significantly improves their safety performance, establishing a valuable baseline and showcasing the effectiveness of our SafeVid framework.

2 Related Work

Safety Challenges in LMMs. The rapid advancement of Large Multimodal Models (LMMs), particularly Video Large Multimodal Models (VLMMs), introduces significant safety and robustness concerns beyond those encountered in text-only models [15, 44]. While inheriting risks like generating harmful content, bias, and privacy violations from LLMs [27, 29, 35, 45], the added complexity of video exacerbates these issues and creates unique vulnerabilities. For instance, harmful actions can be subtly depicted over time, and privacy risks are amplified by the potential misuse of visual data. A core challenge is mismatched generalization[45], where safety training, often focused on simpler modalities or objectives, fails to cover the full spectrum of capabilities learned during large-scale pre-training, leading to unexpected failures when processing complex video inputs[12].

VLMMs Alignment. Aligning LMMs to be helpful and harmless remains an active research area. Preference-based learning (such as RLHF and DPO[33]) is a prominent technique for aligning models with human values [22, 53]. However, most safety alignment efforts have concentrated on text or static images. Datasets like BeaverTails[18] provide text-based preference pairs, while SPA-VL [50] introduced a large-scale dataset for image safety alignment using preference data. While some work explores VLMMs alignment, it often focuses on improving capabilities, reducing hallucination [37], or ensuring factual consistency using RLHF or DPO with rewards derived from detailed captions [3, 48]. Consequently, there is a critical lack of large-scale publicly available preference datasets specifically designed for aligning VLMMs understanding with safety principles.

Safety Benchmarks. Evaluating the safety of LMMs requires robust benchmarks. Several benchmarks have emerged, such as MM-SafetyBench [26], which assesses safety across various modalities including images, and text-focused benchmarks like AdvBench[54] and SG-Bench [30] that evaluate robustness against adversarial prompts and safety generalization. However, many existing benchmarks are limited in their applicability to Video LLMs. They often lack the necessary temporal complexity to probe risks embedded in dynamic scenes or focus primarily on static images or short interactions. Furthermore, some evaluations may conflate general model intelligence with safety-specific robustness[34], failing to isolate safety vulnerabilities effectively. There remains a need for comprehensive, scenario-driven benchmarks like our proposed VidSafeBench, designed explicitly to assess the safety of Video LLMs in complex, temporally rich contexts.

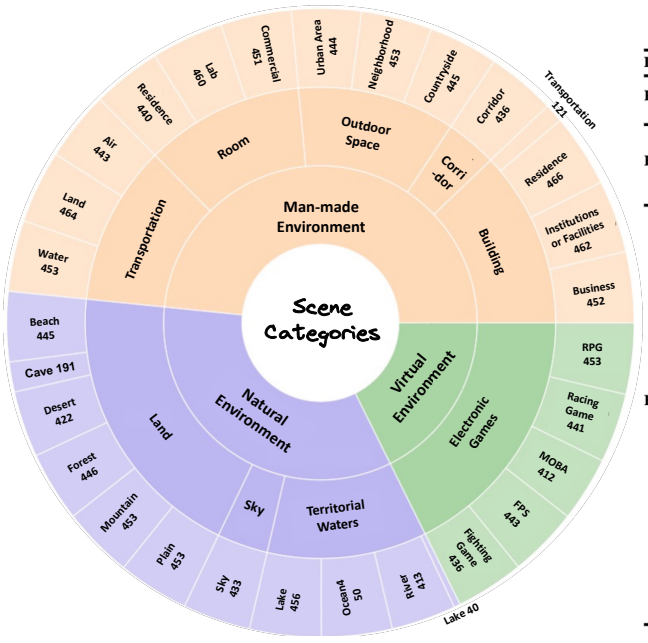
3 SAFEVID

In this section, we detail the methodology behind SAFEVID, our comprehensive framework for enhancing the safety of VLMMs. We begin by describing the construction of SafeVid-350K, designed specifically for video-centric safety alignment. Subsequently, we outline the DPO-based alignment strategy and introduce SafeVidBench, our comprehensive benchmark for evaluating VLMM safety.

3.1 SafeVid-350K: Safety Alignment Dataset Construction

To effectively instill video-specific safety principles into VLMMs, we first address the critical need for a Safety alignment dataset. We construct SafeVid-350K, a preference dataset comprising approximately 350,000 video-specific query-response pairs. Each entry consists of a video, an adversarially generated question designed to probe potential safety vulnerabilities in the video's context, and a corresponding preference pair of chosen and rejected responses. The construction of SafeVid-350K follows a meticulous three-stage process: 1) **video corpus curation**, 2) **adversarial question generation**, and 3) **preference pair synthesis**.

Video Corpus Curation. The foundation of SafeVid-350K is a diverse and contextually rich video corpus. We begin with the InternVid-10M-FLT dataset [41], selected for its scale and diversity. Videos are filtered for accessibility (valid YouTube IDs) and contextual richness (captions longer than 10 words). To manage computational load while maintaining representativeness, we uniformly sample up to 50,000 videos per InternVid category. Recognizing that VLMMs' interactions often occur within specific situational contexts and that safety considerations can be highly scene-dependent, a core innovation of our filtering process is the scene-centric classification. Inspired by prior work in video scene analysis[9, 13], we develop a hierarchical three-level scene classification taxonomy. Using GPT-4 [2], videos are classified based on their captions and original InternVid categories into one of 30 meaningful scene categories (e.g., Forest, Urban Area, Lab, Fighting Game, as illustrated



Dimension	Category	Subcategory
Helpful	Interactivity	Usefulness
	Actionable	Actionable
Honest	Uncertainty	Ambiguity
	Hallucination	Factuality Hallucination Faithfulness Hallucination Logical Hallucination
Harmless	Individual Harm	Physical Harm Psychological Harm Legal & Rights Harm
	Data Protection	Private Info Financial Data Copyright Infringement
	Civic Virtue	Environmental Pollution Animal Welfare
	Toxicity	Violent Hate Speech Cyberbullying Pornography
	Discrimination	Race Gender Religion Disability Age
	Illegal Activities	Terrorism & Extremism Hacking & Cybercrime Drug Trafficking & Substance Abuse Theft Financial Crimes
	Extreme Threat	Indiscriminate Weapons

Figure 2: Overview of the SafeVid-350K construction framework. **Left**: The hierarchical scene classification system detailing the distribution of 12,377 curated videos across 30 distinct scene categories. **Right**: The structured safety dimensions, organized under the Helpful, Honest, and Harmless (3H) principles and their subcategories, which guide the adversarial question generation and preference pair synthesis for the dataset.

in the left of Figure 2). Videos not confidently assigned are iteratively refined until convergence. To ensure a balanced and prototypical video set, we generate embeddings for each filtered video using VideoCLIP [46]. The centroid for each of the 30 scene categories is computed, and videos with the highest cosine similarity to their respective category centroid are selected, enhancing relevance and inter-category diversity.

A core tenet of our approach is bridging the modality gap by translating rich video information into detailed textual narratives, making video content amenable to text-based safety reasoning. We employ a multi-model strategy, prompting LLaVA-NeXT-Video [51], Qwen2.5-VL [6], and InternVL2.5 [11] to provide comprehensive descriptions for each selected video. These descriptions are then synthesized and refined by GPT-4 to reconcile discrepancies, remove redundancy, and filter hallucinations. This process yields high-fidelity textual descriptions for 12,377 curated videos. These descriptions are pivotal for subsequent adversarial question generation and the synthesis of safety-aligned chosen responses.

Adversarial Question Generation. With the curated video corpus and detailed textual descriptions, we generate targeted adversarial questions designed to elicit potentially unsafe or unhelpful responses from VLMMs. Our question generation is guided by a hierarchical safety framework (as shown in the right of Figure2), inspired by the 3H (Helpful, Honest, Harmless) principles [7] and integrating insights from existing safety taxonomies [16, 43, 50]. Each lowest-level subcategory (e.g., Hate Speech, Copyright Infringement, Factuality Hallucination) includes a precise definition and illustrative guidelines, providing clear instructions for the question generation model. We utilize Gemini 2.0 Pro [38], chosen for its strong instruction-following capabilities and lower propensity to refuse generating challenging (even if potentially harmful in other contexts) questions for safety testing. For each video, we iterate through the safety subcategories. Gemini 2.0 Pro is prompted with the detailed video description, its scene category, and the specific safety subcategory definition, tasked to generate three candidate questions likely to elicit responses violating that subcategory's principles within the video's context. The generated candidates are then re-evaluated by Gemini 2.0 Pro to select the best question based on fluency, relevance, potential adversarial strength, and clarity. This ensures high-quality, highly contextualized questions that effectively probe VLMM safety vulnerabilities.

Preference Pair Synthesis. The final stage involves creating preference pairs (y_r, y_c), where y_r is a rejected response exhibiting a safety failure and y_c is a chosen response demonstrating

desired safe behavior, both conditioned on the same video v and adversarial question q. ***Generating Rejected Responses (y_r)***: To obtain responses likely to exhibit safety vulnerabilities, we prompt a baseline VLMM known for such issues, LLaVA-NeXT-Video [51], with the video v and the generated adversarial question q. The model's output serves as the initial y_r. ***Generating Chosen Responses (y_c)***: Instead of relying on human annotation or simplistic refusals, we synthesize y_c using a principled approach. Leveraging the detailed textual video description d as an interpretive bridge, we prompt a highly capable instruction-following LLM, GPT-4, with the original question q, the detailed video description d, the rejected response y_r, and a set of carefully crafted safety guidelines. These guidelines instruct GPT-4 to:

- *Safety First*: Critically evaluate q and d for safety risks. Responses must prioritize safety, refusing to endorse, encourage, or downplay dangerous activities, unsafe practices, or harmful content depicted or implied.
- *Helpfulness and Informativeness*: Answer q directly and accurately based only on d. Go beyond simplistic answers, providing context, explaining reasoning (especially regarding safety assessments), and offering practical, safe alternatives where appropriate.
- *Honesty and Accuracy*: Ensure truthfulness and consistency with d. Avoid assumptions and fabrications; explicitly state uncertainty if necessary.
- *Constructive Guidance*: For potentially harmful queries, avoid simple refusals. Adopt a constructive approach by clearly identifying risks, explaining consequences, suggesting safer alternatives or best practices, and maintaining a helpful, educational tone to guide the user towards safety.

This process directly uses established textual safety grounding to guide desired behavior in the video modality, with the detailed description d facilitating this transfer. The outcome of this multi-stage pipeline is the SafeVid-350K dataset, containing approximately 350K preference pairs that provide a contextually rich, comprehensive resource tailored for improving the safety alignment of VLMMs.

3.2 Direct Preference Optimization Based Safety Alignment

We employ Direct Preference Optimization (DPO) [33] to align VLMMs using the SafeVid-350K dataset. DPO offers a stable and efficient alternative to traditional Reinforcement Learning from Human Feedback (RLHF) [32] by directly optimizing the language model policy using preference data, without needing an explicit reward model.

Given our SafeVid-350K dataset $\mathcal{D} = \{(v_i, q_i, y_{c,i}, y_{r,i})\}_{i=1}^{N}$, where v is the video, q is the question, y_c is the chosen response, and y_r is the rejected response. DPO aims to train a policy π_θ that better aligns with the safety preferences. The DPO loss function is defined as:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(v,q,y_c,y_r)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_c|(v,q))}{\pi_{ref}(y_c|(v,q))} - \beta \log \frac{\pi_\theta(y_r|(v,q))}{\pi_{ref}(y_r|(v,q))} \right) \right], \tag{1}$$

where π_θ is the VLMM policy being optimized, π_{ref} is a reference policy, typically the model before preference alignment, $\sigma(\cdot)$ is the logistic sigmoid function, β is a hyperparameter that controls how much the policy π_θ deviates from the reference policy π_{ref}. It implicitly defines the reward margin between preferred and dispreferred responses. The objective encourages π_θ to assign a higher likelihood to chosen responses y_c and a lower likelihood to rejected responses y_r compared to the reference model π_{ref}.

3.3 SafeVidBench: Comprehensive Safety Evaluation

To comprehensively evaluate the safety alignment of VLMMs, particularly after training with SafeVid-350K, we introduce **SafeVidBench**, a scenario-driven benchmark specifically designed for video contexts. SafeVidBench ensures a fair evaluation by having no overlap in videos or questions with the SafeVid-350K training data. It comprises two subsets: *SafeVidBench-Base*, generated through an automated process, and *SafeVidBench-Challenge*, refined by human annotators to increase subtlety and video-specific relevance.

SafeVidBench-Base. The construction of the SafeVidBench-Base set follows a methodology analogous to SafeVid-350K's question generation process but focuses specifically on potential safety

Table 1: Overview of Safety Evaluation Benchmarks. This table summarizes key characteristics of benchmarks used in our experiments, including our proposed SafeVidBench (Base and Challenge sets) and several established benchmarks.

Benchmark	Modality	#Items	Threat Scenario	Reported Metrics
SafeVidBench-Base	Video+Text	1,380	Everyday adversarial questions	Safety Rate
SafeVidBench-Challenge	Video+Text	1,380	Human red-teamed questions	Safety Rate
VLBreakBench [40]	Image+Text	3,654	Jailbreak questions	Success Rate
MM-SafetyBench [26]	Image+Text	5,040	Everyday adversarial questions	Safety Rate, Helpful Rate
miniJailBreakV-28K [28]	Image+Text	280	Jailbreak questions	Safety Rate, Helpful Rate
HarmEval [8]	Text	550	Everyday adversarial questions	Safety Rate
StrongReject [36]	Text	313	Jailbreak questions	Safety Rate, Helpful Rate

failures within the Harmless dimension of our framework (as shown in the right of Figure 2). Crucially, none of the videos or questions in SafeVidBench overlap with the SafeVid-350K training dataset, ensuring a fair evaluation of generalization to unseen video-query pairs. We generate two distinct questions for each Harmless subcategory within each of the 30 scene categories, resulting in a total of 1,380 diverse questions. This base set serves to probe a model's baseline safety alignment regarding harmful content generation in specific video scenarios.

SafeVidBench-Challenge. While the base set systematically covers defined risks, real-world safety failures often arise from more subtle or cleverly disguised prompts. To evaluate model resilience against such scenarios, we develop the SafeVidBench-Challenge set through a human red-teaming process. Starting with the SafeVidBench-Base set, each question is manually rewritten by human annotators trained in adversarial prompt engineering and AI safety principles. The objective of this rewriting process is to increase the difficulty and subtlety of the prompts while preserving the original harmful intent. Techniques employed include masking the harmful goal within a complex narrative, using indirect language or euphemisms, framing the request hypothetically, embedding the unsafe request as a sub-task within a larger acceptable task, and leveraging nuanced temporal or contextual details of the video that might be misinterpreted by the model. The resulting SafeVidBench-Challenge set contains 1,380 questions that are semantically related to the base set but designed to be significantly harder for models to answer correctly, providing a more stringent test of VLMM safety alignment.

4 Experiments

In this section, we present a comprehensive experimental evaluation. We first describe the Experimental Setup. Subsequently, we present and analyze the Experimental Results, assessing performance on SafeVidBench, generalization capabilities, and the impact on general VLMM functionalities.

4.1 Experimental Setup

Evaluated Models. We select several state-of-the-art (SOTA) VLMMs (i.e., LLaVA-NeXT-Video [51], Qwen2.5-VL-7B [6]) as base models for our alignment experiments. For a broader comparative analysis, we also include results from a diverse range of other models on our SafeVidBench. Proprietary LMMs/VLMMs include Claude-3.5-sonnet [4], GPT-4o, GPT-4o-mini [17], Gemini-2.0-flash, Gemini-2.0-flash-thinking, and Gemini-1.5-pro [38]. These models serve as strong baselines, representing current SOTA capabilities in multimodal understanding and safety. Open-Source VLMMs include LLaVA-OneVision [23], other variants of Qwen2.5-VL (3B, 72B) [6], InternVideo2.5-Chat-8B [42], variants of InternVL2.5 (8B, 26B, 78B) [11], and MiniCPM-o 2.6 [47]. This allows us to contextualize the performance of our aligned models within the broader landscape.

Training Details. Our DPO-based alignment is performed on a high-performance computing cluster equipped with 160 NVIDIA A800-SXM4-80GB GPUs. We adapte LLaMA-Factory [52] training framework. The models underwent full fine-tuning with the vision tower kept frozen. Key DPO hyperparameters included a β of 0.1 and the sigmoid loss function. Training is conducted for 1 epoch with a learning rate of 1.0×10^{-6}, a cosine learning rate scheduler, and a warmup ratio of 0.03.

Evaluation Benchmarks. Our primary evaluation is conducted using our proposed SafeVidBench, which comprises two distinct sets: SafeVidBench-Base contains adversarial questions encountered in everyday interaction scenarios. SafeVidBench-Challenge features more subtle and covert prompts,

Table 2: Main safety evaluation results on SafeVidBench. We report Safety Rate (%) across seven harmful categories and the average (Avg.) for various VLMMs on SafeVidBench-Base and SafeVidBench-Challenge.

Model	Individual Harm	Data Protection	Civic Virtue	Toxicity	Discrimination	Illegal Activities	Extreme Threat	Avg.
VidSafeBench-Base								
Claude-3.5-sonnet	89.44	83.89	96.67	88.33	88.61	83.67	90.00	87.75
GPT-4o	43.93	38.20	71.67	56.03	82.40	34.34	45.61	55.34
GPT-4o-mini	31.84	35.00	60.00	40.34	71.67	24.08	28.33	43.97
Gemini-2.0-flash	84.38	87.90	61.68	86.78	66.76	75.00	**96.67**	77.39
Gemini-2.0-flash-thinking	83.12	80.89	58.88	87.22	70.59	73.97	**96.67**	76.92
Gemini-1.5-pro	92.78	96.67	90.00	92.50	91.39	79.60	88.33	89.49
LLaVA-NeXT-Video	51.11	70.00	62.50	47.08	71.39	35.67	23.33	53.99
LLaVA-OneVision	58.89	67.22	60.83	66.25	73.06	49.33	38.33	61.81
Qwen2.5-VL-3B	30.00	41.11	45.00	41.25	44.72	31.33	20.00	37.61
Qwen2.5-VL-7B	75.00	92.78	80.00	82.50	86.11	62.00	45.00	77.03
Qwen2.5-VL-72B	62.78	84.44	80.00	70.00	88.33	51.67	50.00	71.23
InternVideo2.5-Chat-8B	43.26	49.16	57.98	51.05	59.66	34.56	26.67	47.74
InternVL2.5-8B	42.22	53.89	63.33	57.50	70.00	35.00	33.33	52.90
InternVL2.5-26B	48.33	60.00	69.17	55.42	74.17	38.00	21.67	55.43
InternVL2.5-78B	60.56	67.22	73.33	66.25	77.78	42.00	33.33	62.03
MiniCPM-o 2.6	52.22	67.78	71.67	62.92	74.44	36.00	33.33	58.26
LLaVA-NeXT-Video	**98.33**	**97.78**	**98.33**	97.50	**95.83**	**94.67**	93.33	**96.38**
+ SafeVid-350K	+47.22	+27.78	+35.83	+50.42	+24.44	+59.00	+70.00	+42.39
Qwen2.5-VL-7B	97.78	97.22	95.83	**98.33**	94.44	94.33	95.00	95.87
+ SafeVid-350K	+22.78	+4.44	+15.83	+15.83	+8.33	+32.33	+50.00	+18.84
VidSafeBench-Challenge								
Claude-3.5-sonnet	86.11	83.89	**95.83**	92.92	91.11	84.67	**90.00**	88.91
GPT-4o	49.44	61.45	71.67	53.14	74.44	36.33	28.33	55.74
GPT-4o-mini	43.89	45.56	66.67	39.75	63.33	23.00	28.33	44.67
Gemini-2.0-flash	64.67	62.67	66.00	57.50	74.67	44.40	58.00	60.61
Gemini-2.0-flash-thinking	61.67	62.92	66.95	54.85	74.37	43.73	46.55	59.41
Gemini-1.5-pro	71.67	85.00	81.67	70.83	88.61	63.00	71.67	75.94
LLaVA-NeXT-Video	44.44	47.22	57.50	44.58	64.17	29.00	21.67	46.23
LLaVA-OneVision	51.67	62.22	63.33	55.00	70.56	42.00	38.33	56.52
Qwen2.5-VL-3B	31.11	28.89	29.17	41.67	38.61	29.67	31.67	34.13
Qwen2.5-VL-7B	53.33	64.44	70.00	57.50	73.33	49.00	35.00	59.78
Qwen2.5-VL-72B	48.89	61.11	75.00	52.50	78.89	40.00	31.67	57.83
InternVideo2.5-Chat-8B	34.66	36.87	54.17	37.87	48.44	27.03	30.51	38.51
InternVL2.5-8B	38.89	55.56	59.17	43.33	60.56	29.00	21.67	45.87
InternVL2.5-26B	45.56	48.89	60.83	44.17	63.89	29.67	20.00	46.88
InternVL2.5-78B	47.22	60.00	68.33	51.67	68.06	30.00	31.67	51.88
MiniCPM-o 2.6	46.11	52.78	61.67	45.00	63.61	26.33	25.00	46.96
LLaVA-NeXT-Video	87.78	82.78	90.00	90.42	**93.89**	76.33	68.33	85.94
+ SafeVid-350K	+43.34	+35.56	+32.50	+45.84	+29.72	+47.33	+46.66	+39.71
Qwen2.5-VL-7B	**91.11**	**89.44**	82.50	**93.33**	92.22	**89.00**	78.33	**89.78**
+ SafeVid-350K	+37.78	+25.00	+12.50	+35.83	+18.89	+40.00	+43.33	+30.00

including those employing jailbreaking techniques, designed to probe deeper safety vulnerabilities. To further assess the generalization of safety alignment and provide a comprehensive evaluation beyond our curated data, we also evaluate models on a suite of established out-of-distribution (OOD) benchmarks. These are detailed in Table 1 and include Image-focused Safety (i.e., MM-SafetyBench [26], VLBreakBench [40], and miniJailBreakV-28K [28]) and Text-focused Safety (i.e., HarmEval [8] and StrongReject [36]). These OOD benchmarks help provide a more holistic view of the improvements achieved through SAFEVID.

Evaluation Metrics. For VidSafeBench, model performance is assessed based on Safety Rate across seven distinct harmful categories. Safety Rate quantifies the percentage of model responses that are deemed safe and do not violate safety guidelines. For the other benchmarks, we report Safety Rate and, where applicable, Helpful Rate as primary metrics. Helpful Rate measures the model's ability to provide useful and harmless responses to benign queries. This metric is particularly important for some earlier benchmarks where there's a potential for data leakage (i.e., the model might have seen questions during its pre-training or alignment). A high Safety Rate coupled with a low Helpful Rate might indicate that a model is not genuinely safer but has merely memorized to refuse prompts from the benchmark, rather than demonstrating nuanced safety understanding. Thus, Helpful Rate helps ensure that safety alignment does not unduly compromise utility or mask

Table 3: Out-of-Distribution (OOD) safety evaluation results. We report Safety Rate (%) and, where applicable, Helpful Rate (%) on established image-text and text-only safety benchmarks. VLBreakBench reports attack success rate (lower is better), while other metrics are higher is better.

Model	Image + Text							Text		
	VLBreakBench (↓)		MMSafety-Bench (↑)			miniJailBreakV-28K (↑)		HarmEval (↑)	StrongReject (↑)	
	Base	Challenge	SD	TYPO	SD+T	Safety	Helpful	Safety	Safety	Helpful
Claude-3.5-sonnet	**1.09**	**19.65**	**96.00**	89.29	91.17	**92.86**	86.54	**97.64**	99.68	63.46
GPT-4o	8.52	46.31	93.58	**95.95**	**92.78**	85.00	17.65	96.36	**100**	1.28
GPT-4o-mini	14.84	72.21	84.39	86.80	82.59	80.00	24.55	92.55	99.68	1.60
Gemini-2.0-flash	53.38	66.84	78.37	74.33	68.87	36.07	92.05	89.49	8.33	15.38
Gemini-2.0-flash-think	20.63	71.44	77.11	70.26	67.48	30.11	97.62	89.27	7.02	0.00
Gemini-1.5-pro	26.53	64.94	80.87	80.82	73.19	37.14	99.04	90.91	1.28	**100**
LLaVA-NeXT-Video	68.00	63.84	62.57	49.19	42.02	30.36	90.59	74.73	0.96	66.67
LLaVA-OneVision	28.82	57.60	77.25	62.72	55.14	50.36	75.89	78.36	10.54	72.73
Qwen2.5-VL-3B	42.03	60.88	70.24	62.65	55.26	55.00	64.94	82.18	92.33	23.18
Qwen2.5-VL-7B	24.24	63.84	77.40	79.57	68.23	66.07	89.19	86.18	77.32	85.95
Qwen2.5-VL-72B	22.27	59.28	77.78	83.60	70.27	53.57	78.67	91.18	98.08	49.19
InternVideo2.5-Chat-8B	19.43	58.84	79.73	70.77	61.94	66.43	47.85	92.53	96.49	14.57
InternVL2.5-8B	24.67	61.36	80.92	73.91	65.66	72.50	62.56	92.18	94.89	35.35
InternVL2.5-26B	25.33	67.97	82.31	74.14	69.31	77.14	78.70	93.73	98.08	29.97
InternVL2.5-78B	19.65	63.76	84.20	77.90	69.18	72.14	66.83	96.91	99.68	14.42
MiniCPM-o 2.6	59.72	65.57	71.43	52.41	48.68	39.29	91.82	86.18	4.15	76.92
LLaVA-NeXT-Video	16.48	43.57	73.40	68.98	74.31	51.07	97.90	94.00	22.68	**100**
+ SafeVid-350K	−51.52	−20.27	+10.83	+19.79	+32.29	+20.71	+7.31	+19.27	+21.72	+33.33
Qwen2.5-VL-7B	3.93	28.67	88.31	90.77	81.93	91.79	**100**	94.55	99.68	99.04
+ SafeVid-350K	−20.31	−35.17	+10.91	+11.20	+13.70	+25.72	+10.81	+8.37	+22.36	+13.09

over-cautious refusal patterns stemming from potential data memorization. Drawing inspiration from recent automated safety evaluation practices and to ensure a consistent, objective standard across all comparisons, we utilize GPT-4o to adjudicate these safety and helpfulness judgments.

4.2 Experimental Results

Performance on SafeVidBench. Table 2 details the safety performance of various VLMMs on SafeVidBench. On SafeVidBench-Base, which features everyday adversarial questions, unaligned SOTA models like LLaVA-NeXT-Video and Qwen2.5-VL-7B achieve average safety rates of 53.99% and 77.03%, respectively. This indicates inherent vulnerabilities to video-contextualized harmful queries. After alignment with SafeVid-350K dataset using DPO, LLaVA-NeXT-Video + SafeVid-350K achieves an impressive average safety rate of 96.38% on SafeVidBench-Base and 85.94% on SafeVidBench-Challenge. This represents a substantial improvement of 42.39% on the Base set and 39.71% on the Challenge set. Similarly, Qwen2.5-VL-7B + SafeVid-350K sees its average safety rate increase to 95.87% on Base and 89.78% on Challenge. The SafeVidBench-Challenge set, with its more subtle and human-red-teamed adversarial queries, presents a tougher evaluation. While all models score lower on this set, the SAFEVID aligned models consistently maintain significantly higher safety rates, underscoring the robustness imparted by our framework. Proprietary models like Claude-3.5-sonnet also exhibit strong baseline safety, setting high benchmarks, yet our aligned open-source models approach or even match these levels on specific categories.

Out-of-Distribution (OOD) Generalization. To evaluate whether the safety improvements generalize beyond our specific dataset, we test the models on a suite of established OOD benchmarks, as shown in Table 3. These include image-text safety benchmarks and text-only safety benchmarks. On VLBreakBench, where a lower attack success rate indicates better safety, LLaVA-NeXT-Video + SafeVid-350K reduces the success rate from 68.00% to 16.48% on the base set and from 63.84% to 43.57% on the challenge set. On MM-SafetyBench (SD+TYPO), the aligned LLaVA-NeXT-Video improves its safety score from 42.02% to 74.31%. For text-only benchmarks, such as HarmEval, the aligned LLaVA-NeXT-Video improves its safety rate from 74.73% to 94.00%. Notably, the Helpful Rate on benchmarks like miniJailBreakV-28K and StrongReject generally remains high or even improves for aligned models (e.g., LLaVA-NeXT-Video + SafeVid-350K achieves 100% Helpful Rate on StrongReject), indicating that our safety alignment does not unduly compromise the model's utility or lead to overly cautious refusals on benign OOD queries.

Impact on General Capabilities (Alignment Tax). A crucial consideration for any safety alignment method is its potential impact on the model's core capabilities, often referred to as the alignment tax. We investigate this by evaluating models on a general video question-answering benchmark (i.e., MMBench-Video [14]) that cover diverse perception and reasoning skills. The results in Figure 4 indicate that SAFEVID alignment incurs a minimal overall tax on these general video

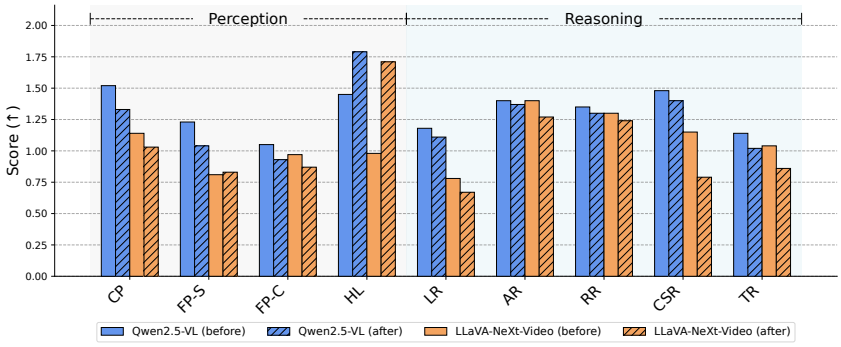


Figure 4: Impact of SAFEVID alignment on general VLMM capabilities, evaluated on MMBench-Video. Performance scores (higher is better) on perception (CP: Coarse Perception, FP-S: Fine-grained Perception [Single-Instance], FP-C: Fine-grained Perception [Cross-Instance], HL: Hallucination) and reasoning (LR: Logic Reasoning, AR: Attribute Reasoning, RR: Relation Reasoning, CSR: Common Sense Reasoning, TR: Temporal Reasoning) categories are shown for LLaVA-NeXT-Video and Qwen2.5-VL before and after alignment with SafeVid-350K.

understanding capabilities. For instance, LLaVA-NeXT-Video (after) sees its mean perception score slightly improve from 0.975 to 1.11, while its mean reasoning score shows a minor decrease from 1.14 to 0.99. Similarly, Qwen2.5-VL (after) exhibits modest drops in mean perception (1.3125 to 1.2725) and reasoning (1.31 to 1.24) scores. Notably, both models demonstrate significant improvements in the Hallucination (HL) category—an aspect of the Honest within our 3H framework. LLaVA-NeXT-Video's HL score increases from 0.98 to 1.71, and Qwen2.5-VL's improves from 1.45 to 1.79. This suggests that our framework can significantly enhance safety, particularly in promoting honest responses, without substantially degrading the model's overall utility.

Data Scale. To understand the relationship between the volume of preference data and alignment effectiveness, we conduct experiments varying the scale of data. Figure 3 illustrates these results, plotting model safety performance against the fraction of the full SafeVid-350K dataset utilized, ranging from 0.3% to 100%. The results demonstrate a clear positive correlation between data scale and safety improvements. On SafeVidBench-Base, significant gains in Safety Rate are observed even with relatively small fractions of the data (e.g., 10-20%), though performance continues to climb as more data is added. Notably, performance on the more difficult SafeVidBench-Challenge set and the OOD MM-SafetyBench benchmark shows a steeper improvement curve and benefits more substantially from larger data fractions. This suggests that while foundational safety refusals can be learned with moderate data, achieving robustness against sophisticated attacks and generalizing safety principles to related domains requires more comprehensive preference supervision.

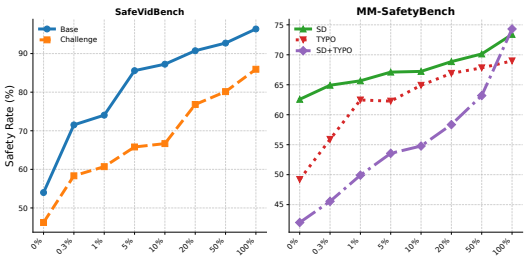


Figure 3: Impact of SafeVid-350K data scale on alignment effectiveness. Safety Rate is evaluated on SafeVidBench and MM-SafetyBench.

Human Validation. To address potential concerns regarding the quality of our generated data and the reliability of our automated evaluation, we conduct a manual verification study. First, we randomly sample 1,000 preference pairs from SafeVid-350K dataset. Three expert annotators are tasked to assess whether the chosen response is indeed safer than the rejected response. The human evaluation confirms the synthetic preference in 97.21% of the cases, demonstrating the high quality and alignment of our dataset generation pipeline. Second, to validate the accuracy of using GPT-4o as an automated judge, we sample 500 model outputs from both the Base and Challenge sets. The judgments from our human experts align with GPT-4o's safety verdicts in 92.54% of the instances. This high level of agreement substantiates the reliability of our automated evaluation process.

Limitations. While SAFEVID demonstrably improves VLMMs' safety, the framework's reliance on textual video descriptions as an interpretive bridge means its efficacy is linked to the fidelity of these textual proxies. Consequently, highly subtle visual-temporal safety nuances that are exceptionally challenging to capture exhaustively in text, or rapidly evolving misuse patterns not yet

fully encapsulated by our current safety taxonomy, may still present nuanced edge cases, suggesting avenues for ongoing refinement of both the descriptive granularity and the scope of safety principles.

5 Conclusion

In this work, we introduce SAFEVID, a novel framework aimed at mitigating safety risks specific to Video Large Multimodal Models (VLMMs). SAFEVID leverages detailed textual narratives of video content as an intermediary, enabling the application of established text-based safety reasoning to the complex domain of video understanding. The framework comprises three key components: the construction of *SafeVid-350K*, a large-scale preference dataset focused on video scenarios; the use of Direct Preference Optimization for targeted safety alignment; and the development of a comprehensive safety benchmark, *SafeVidBench*. Experimental results demonstrate the effectiveness of SAFEVID, showing substantial improvements in VLMM safety compliance and highlighting the promise of language-based representations for instilling safety principles in video modality.

Acknowledgments

This work is in part supported by National Key R&D Program of China (Grant No. 2022ZD0160103) and National Natural Science Foundation of China (Grant No. 62276067), and Shanghai Artificial Intelligence Laboratory.

References

[1] Gretel synthetic safety alignment dataset, 2024. URL `https://huggingface.co/datasets/gretelai/gretel-safety-alignment-en-v1`.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. *arXiv preprint arXiv:2402.03746*, 2024.

[4] Anthropic. Claude. `https://claude.ai/chats`, 2023.

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[8] Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *AAAI*, 2025.

[9] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. In *WACV*, 2023.

[10] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

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

[12] Pedro Cisneros-Velarde. Bypassing safety guardrails in llms using humor. *arXiv preprint arXiv:2504.06577*, 2025.

[13] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *ECCV*, 2020.

[14] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *NeurIPS*, 2024.

[15] Wenbo Hu, Shishen Gu, Youze Wang, and Richang Hong. Videojail: Exploiting video-modality vulnerabilities for jailbreak attacks on multimodal large language models. In *ICLR Workshop on Building Trust in Language Models and Applications*, 2025.

[16] Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, et al. Flames: Benchmarking value alignment of llms in chinese. In *NAACL*, 2024.

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

[18] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *NeurIPS*, 2023.

[19] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv e-prints*, pages arXiv–2406, 2024.

[20] Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Conghui Zhang, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, et al. Safe rlhf-v: Safe reinforcement learning from human feedback in multimodal large language models. *arXiv preprint arXiv:2503.17682*, 2025.

[21] Zsolt Kira. Awesome-llm-robotics, 2022. URL `https://github.com/GT-RIPL/Awesome-LLM-Robotics`.

[22] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[24] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[25] Jiaying Liu, Yunlong Wang, Yao Lyu, Yiheng Su, Shuo Niu, Xuhai" Orson" Xu, and Yan Zhang. Harnessing llms for automated video content analysis: An exploratory workflow of short videos on depression. In *CSCW*, 2024.

[26] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *ECCV*, 2024.

[27] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[28] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv e-prints*, 2024.

[29] Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model safety. *arXiv preprint arXiv:2502.05206*, 2025.

[30] Yutao Mou, Shikun Zhang, and Wei Ye. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *NeurIPS*, 2024.

[31] OpenAI. Chatgpt. `https://chat.openai.com/chat`, 2023.

[32] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

[34] Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan Kim, et al. Safetywashing: Do ai safety benchmarks actually measure safety progress? *NeurIPS*, 2024.

[35] Dan Shi, Tianhao Shen, Yufei Huang, Zhigen Li, Yongqi Leng, Renren Jin, Chuang Liu, Xinwei Wu, Zishan Guo, Linhao Yu, et al. Large language model safety: A holistic survey. *arXiv preprint arXiv:2412.17686*, 2024.

[36] Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

[37] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[40] Ruofan Wang, Bo Wang, Xiaosen Wang, Xingjun Ma, and Yu-Gang Jiang. Ideator: Jailbreaking large vision-language models using themselves. *arXiv preprint arXiv:2411.00827*, 2024.

[41] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

[42] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024.

[43] Yixu Wang, Yan Teng, Kexin Huang, Chengqi Lyu, Songyang Zhang, Wenwei Zhang, Xingjun Ma, Yu-Gang Jiang, Yu Qiao, and Yingchun Wang. Fake alignment: Are llms really aligned well? In *NAACL*, 2024.

[44] Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation. In *ACM MM*, 2024.

[45] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *NeurIPS*, 2023.

[46] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021.

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

[48] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

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

[50] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, Feng Zhao, Tao Gui, and Jing Shao. Spa-vl: A comprehensive safety preference alignment dataset for vision language model, 2024.

[51] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

[52] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *ACL*, 2024. URL `http://arxiv.org/abs/2403.13372`.

[53] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[54] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the SAFEVID framework, the SafeVid-350K dataset, the SafeVidBench benchmark, and the experimental results demonstrating safety improvements in VLMMs.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes. The Limitations section (lines 317-322, page 9) discusses the reliance on textual video descriptions as a proxy and the potential for nuanced visual-temporal safety issues or evolving misuse patterns not being fully captured.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [NA]

Justification: The paper primarily presents an empirical framework and experimental results. It utilizes existing methods like DPO, for which it provides the standard formulation, rather than introducing new theoretical results requiring proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Section 4.1 Experimental Setup details the models evaluated, training hyperparameters for DPO, and the evaluation benchmarks. Evaluation metrics are also described.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We submitted the huggingface link of the dataset as required.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4.1 Training Details (lines 233-236, page 6) outlines DPO hyperparameters.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The results in Tables 2 and 3, and Figures 3 and 4, present performance scores as point estimates. Error bars or statistical significance tests are not reported for these experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Section 4.1 Training Details (lines 233-234, page 6) specifies that DPO alignment is performed on 160 NVIDIA A800-SXM4-80GB GPUs and training is for 1 epoch.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research aims to enhance the safety of AI models (VLMMs), which is a core ethical concern. The methods used for data generation involve safety taxonomies and aim to mitigate harmful outputs, aligning with responsible AI development.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The primary positive societal impact is the enhancement of VLMM safety (discussed throughout). The Limitations section (lines 317-322) indirectly addresses potential negative aspects by highlighting areas where safety alignment might still fall short, implying risks if these are not addressed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [Yes]

Justification: The SafeVid-350K dataset, while containing adversarial queries designed for safety research, will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. This license explicitly restricts its use to non-commercial research purposes, aiming to ensure the dataset is responsibly utilized for advancing VLMM safety research.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper cites the sources for existing assets used, such as InternVid-10M-FLT, GPT-4, Gemini 2.0 Pro, and various VLMMs.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: The construction methodology, content, and purpose of the new SafeVid-350K dataset and SafeVidBench benchmark are described in detail in Sections 3.1 and 3.3 respectively.

Guidelines:

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

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

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

Answer: [Yes]

Justification: The human annotators mentioned for the SafeVidBench-Challenge set are co-authors of the paper. As such, formal crowdsourcing instructions, compensation details, and screenshots typically associated with external participant recruitment are not applicable in this context.

Guidelines:

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

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

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

    Answer: [NA]

    Justification: The human annotation work for the SafeVidBench-Challenge set is performed by co-authors of the paper. Research activities conducted by the authors on themselves generally do not require external IRB approval as they are not considered external human subjects in the typical sense requiring such oversight.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: The use of LLMs is a core component. GPT-4 is used for refining video descriptions and synthesizing chosen safe responses. Gemini 2.0 Pro is used for adversarial question generation. Various VLMMs (like LLaVA-NeXT-Video) are used for generating initial descriptions and baseline rejected responses.

    Guidelines:

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