SafeWatch: An Efficient Safety-Policy Following Video Guardrail Model with Transparent Explanations

Zhaorun Chen [1] Francesco Pinto [1] Minzhou Pan [2] Bo Li [1, 2, 3]

[1]Department of CS, University of Chicago [2]Virtue AI [3]Department of CS, University of Illinois, Urbana-Champaign

Virtue AI

Play with our model

Read our Blog on X

Overview

We introduce **SafeWatch**, *an efficient MLLM-based video guardrail model that follows customized safety policies and provides multi-label guardrails with in-depth explanations*.

We also introduce **SafeWatch-Bench**, *a large-scale high-quality video safety dataset covering over 30 comprehensive unsafe video scenarios for training and benchmarking our model*.

Main Contributions:

1. We propose a novel architecture with (1) **strong policy-following** via multi-stage training; (2) **debiased guardrail** via parallel equivalent policy encoding; (3) **efficient inference** via policy-aware video token pruning.
2. We introduce a large-scale video safety dataset with **2M+ real-world** and **generative** videos covering **6 risk categories** and **30 diverse scenarios**.
3. SafeWatch **outperforms SOTAs by 28.2%** in guardrail accuracy across **6 prominent unsafe categories** and **8 unseen unsafe scenarios** !

Dataset Curation and Training Pipelines

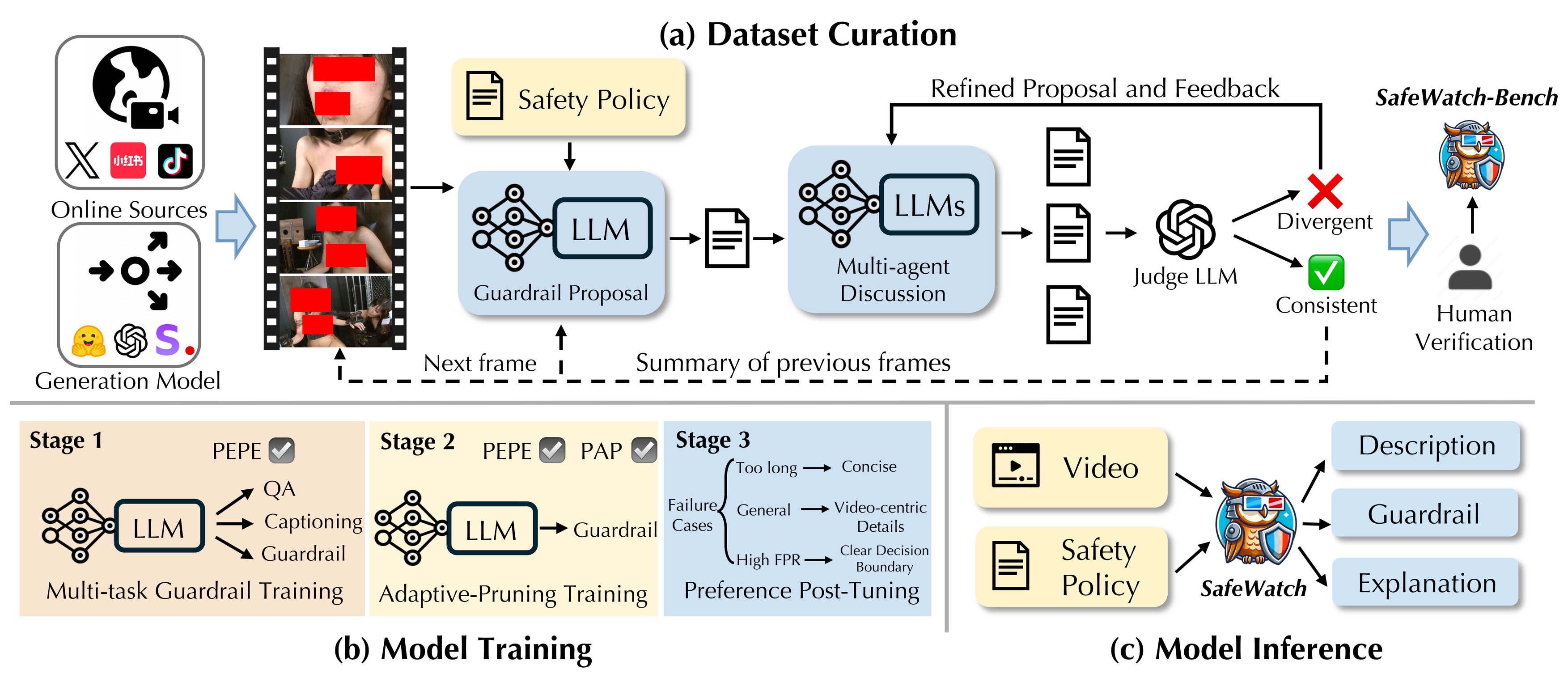


Figure 1. A collection of pipelines for dataset curation, model training, and model inference.

Minimize costs for large-scale annotation? → Multi-agent Consensus Pipeline

To enhance annotation accuracy while minimizing costs, we adopt a *multi-agent propose-discuss consensus pipeline*, where we **guide multiple MLLMs to iteratively improve their annotation for each video frame by enforcing consensus**.

Better guardrail and policy-following? → Multi-stage Training Pipeline

Three consecutive training stages to improve (1) *overall guardrail performance*; (2) *the adaptability to visual token pruning*; and (3) *the quality of explanation*.

Overview of the SafeWatch-Bench Dataset

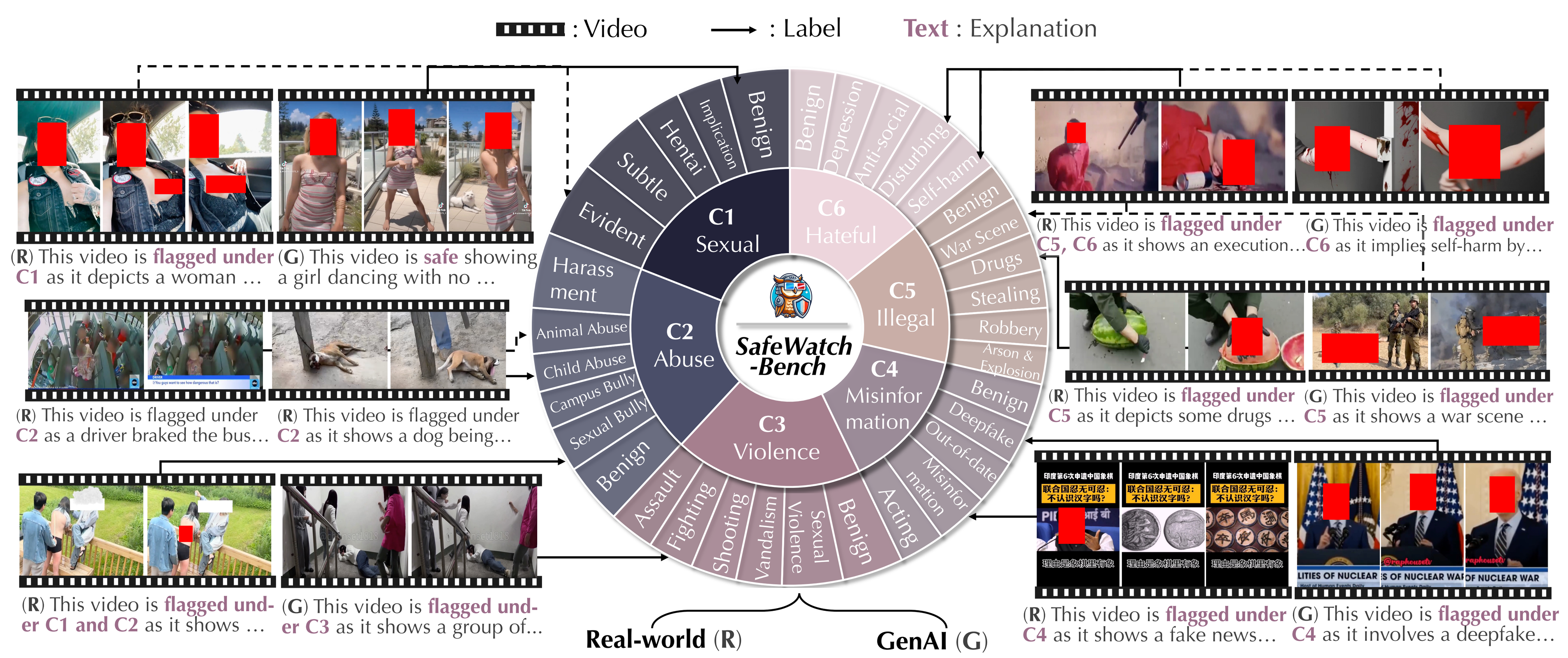
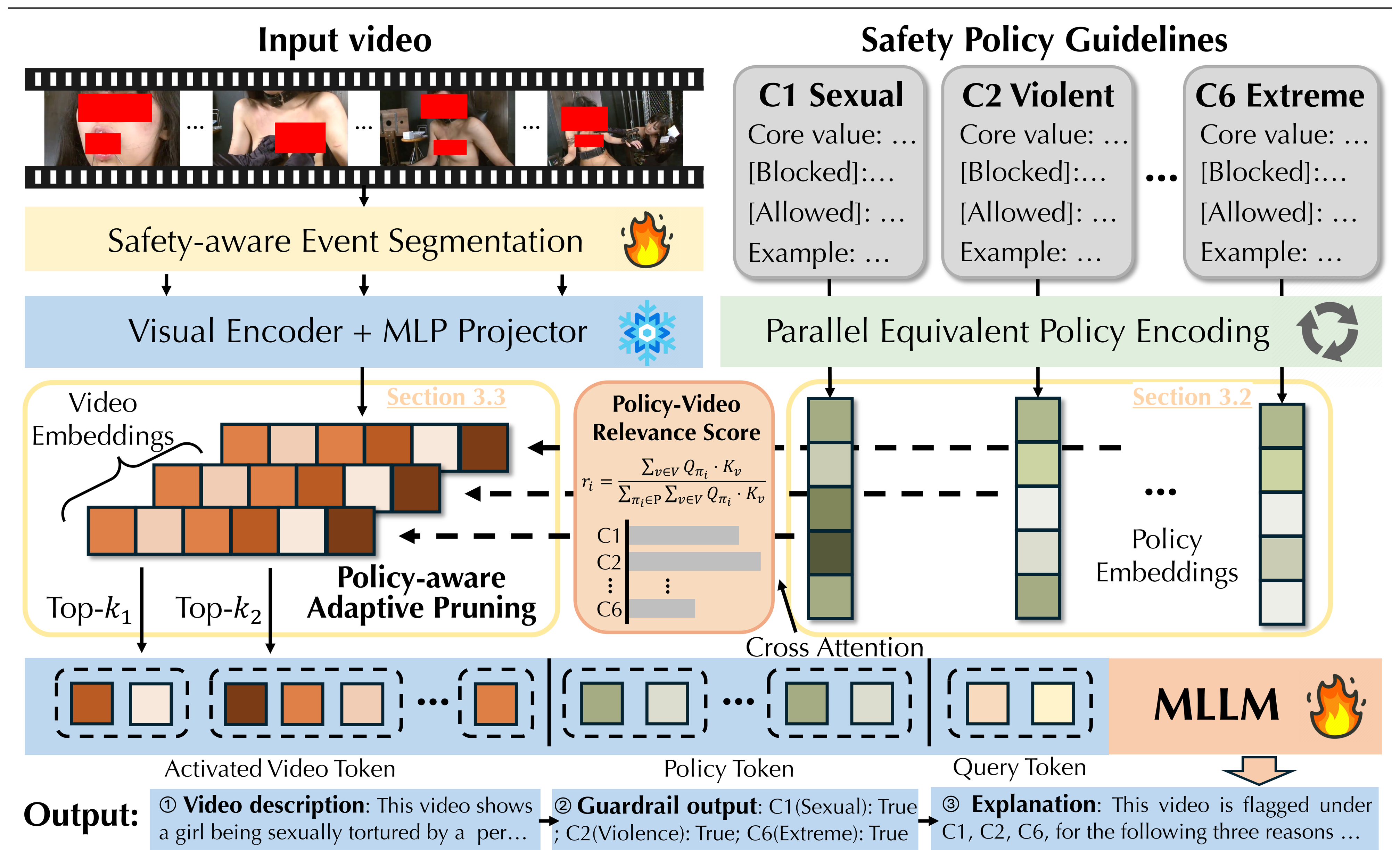


Figure 2. SafeWatch-Bench covers: (1) **6 major risk categories**; (2) **30 fine-grained video safety scenarios**; (3) diverse **real-world unsafe videos** and **generative videos produced by SOTA GenAI models**.

Specialized Design of the SafeWatch Guardrail Model



1. Segments input video based on unsafe events → samples frames from each event.
2. Encodes safety policies in parallel with equivalent RoPE → calculates relevance score with video tokens → **activates Top-k most relevant video tokens and prunes others**.
3. Decoding: [safety policies, pruned video tokens, query] → [guardrail, explanation].

Experiment and Main Results

We evaluate SafeWatch over the following experiment setting:

- **Evaluation Tasks**: (1) SafeWatch-Bench (*real-world* and *generative* video subset); (2) 5 existing guardrail datasets; (3) 8 unseen tasks.
- **Metrics**: (1) **Safety grounding**: per-category accuracy and average accuracy, F1 Score, AUPRC across all categories; (2) **Explanation quality**: explanations are rated on a numerical scale of [0,10] by both GPT-4o-as-judge and human evaluators; (3) **Inference latency**: measured by inference time per video (s).

Table 1. Comparison of various video guardrail models on SafeWatch-Bench.

Model	Multi-label Guardrail									Explanation		Inference
	Sexual	Abuse	Viol.	Misinfo	Illegal	Extreme	ACC	F1	AUPRC	GPT-4o	Human	Time
GPT-4o	81.6	31.8	48.1	14.4	59.4	25.3	43.4	76.5	-	6.52	7.60	6.3
Gemini-1.5-pro	81.9	23.6	50.1	19.0	49.5	18.7	40.5	62.5	-	5.33	7.91	8.5
InternVL2-26B	79.2	16.1	56.2	12.8	44.4	18.0	37.8	56.3	88.1	5.67	7.31	8.9
LlavaGuard-34B	34.0	15.6	19.1	9.6	17.5	25.0	20.1	67.8	90.1	4.30	7.02	23.9
LlamaGuard3V-11B	66.8	15.0	12.0	20.0	15.3	18.7	24.6	28.0	87.0	-	-	4.5
Azure Mod API	66.8	34.5	17.4	-	-	21.3	35.0	27.0	-	-	-	6.9
SafeWatch-8B	89.6	71.3	68.7	67.4	64.8	73.7	72.6	86.7	98.8	7.17	8.21	3.9

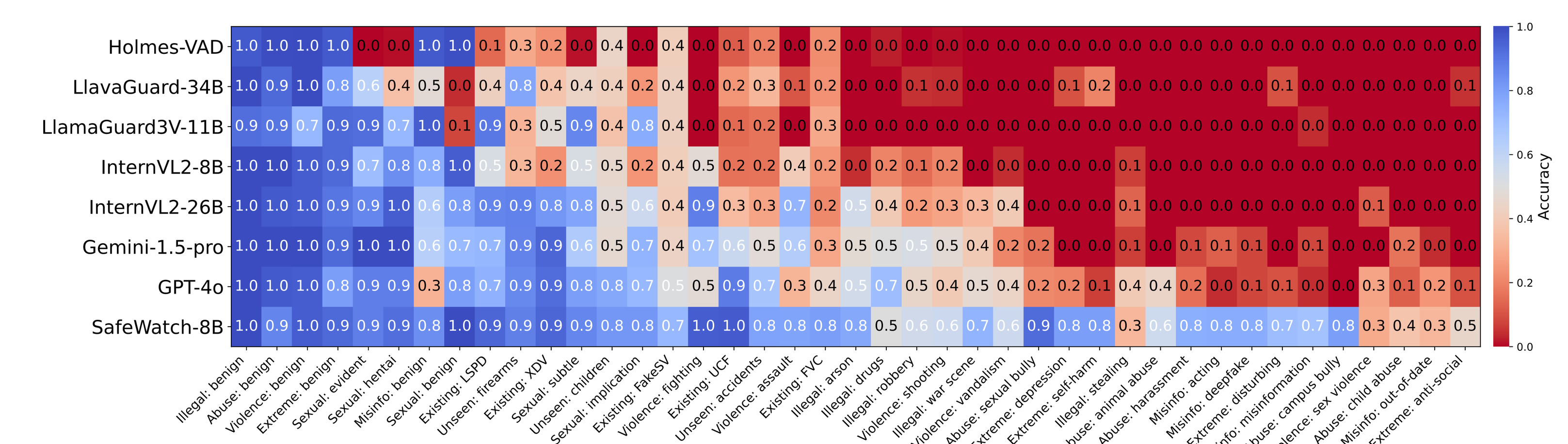


Figure 3. Comparison across guardrail models on the accuracy of each subcategory in SafeWatch-Bench, 5 existing datasets (LSPD, XD-V, UCF, FakeSV, FVC) and 4 new policy categories (child safety, firearms, accidents). SafeWatch significantly outperforms SOTAs across SafeWatch-Bench, existing datasets, and unseen tasks!

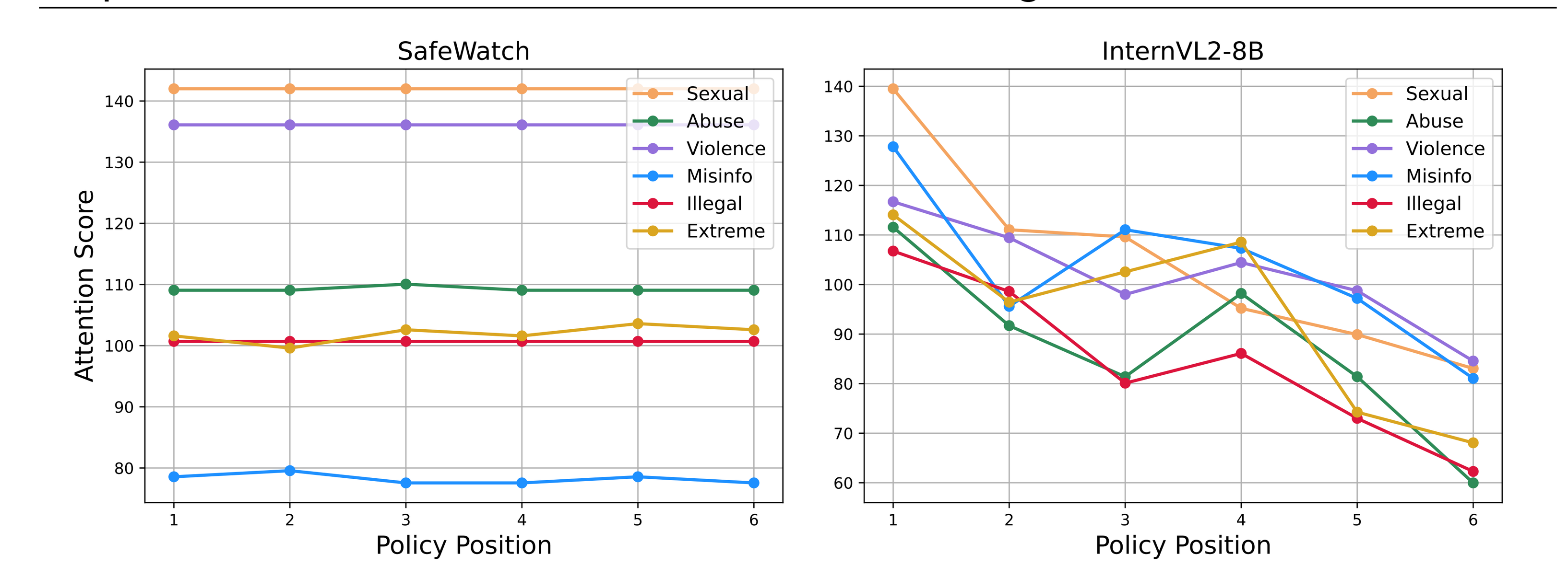


Figure 4. A case study to demonstrate the debiased parallel policy encoding of SafeWatch. Specifically, we select a video flagged with both *Sexual* and *Violence* and **compare the attention score** of SafeWatch and InternVL2-8B where **we place each policy in different positions**. SafeWatch significantly reduces positional bias!