Controllable Multi-label Video Safety Detection via Adaptive Tversky Policy Optimization

Guangyu Yang

University of Cambridge · Machine Intelligence Laboratory

Joint work with Jingbiao Mei, Mingsheng Sun, Jinghong Chen, Yingtong Bu, Pengda Qin, Da Chen and Bill Byrne

Multi-label video safety detection

One video can violate several safety categories.

BINARY DETECTION

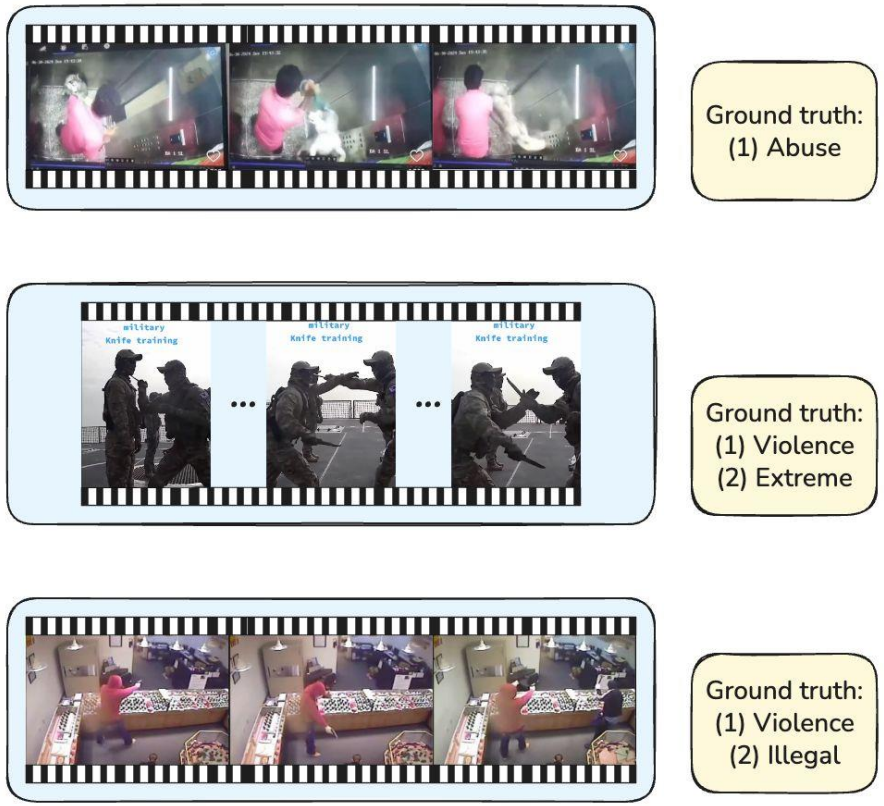
Safe / unsafe

A single decision hides the individual risk categories.

MULTI-LABEL DETECTION

Violence + extremism + …

Predict which unsafe categories are present in each video.



Content warning: potentially disturbing examples.

Moderation needs different precision-recall trade-offs

DIFFERENT SAFETY CATEGORIES

Multi-label safety detection inherently involves heterogeneous risk profiles.

Misinformation — High precision

Prioritize higher precision to reduce over-censorship.

Extreme violence — High recall

Prioritize higher recall to minimize the risk of harmful exposure.

DIFFERENT MODERATION STAGES

Different operating points across the moderation pipeline

Candidate screening — High recall

Capture a broad pool of potentially harmful videos for downstream verification.

Candidate verification — Higher precision

Filter out false alarms to reduce human review. Moderators make the final enforcement decision.

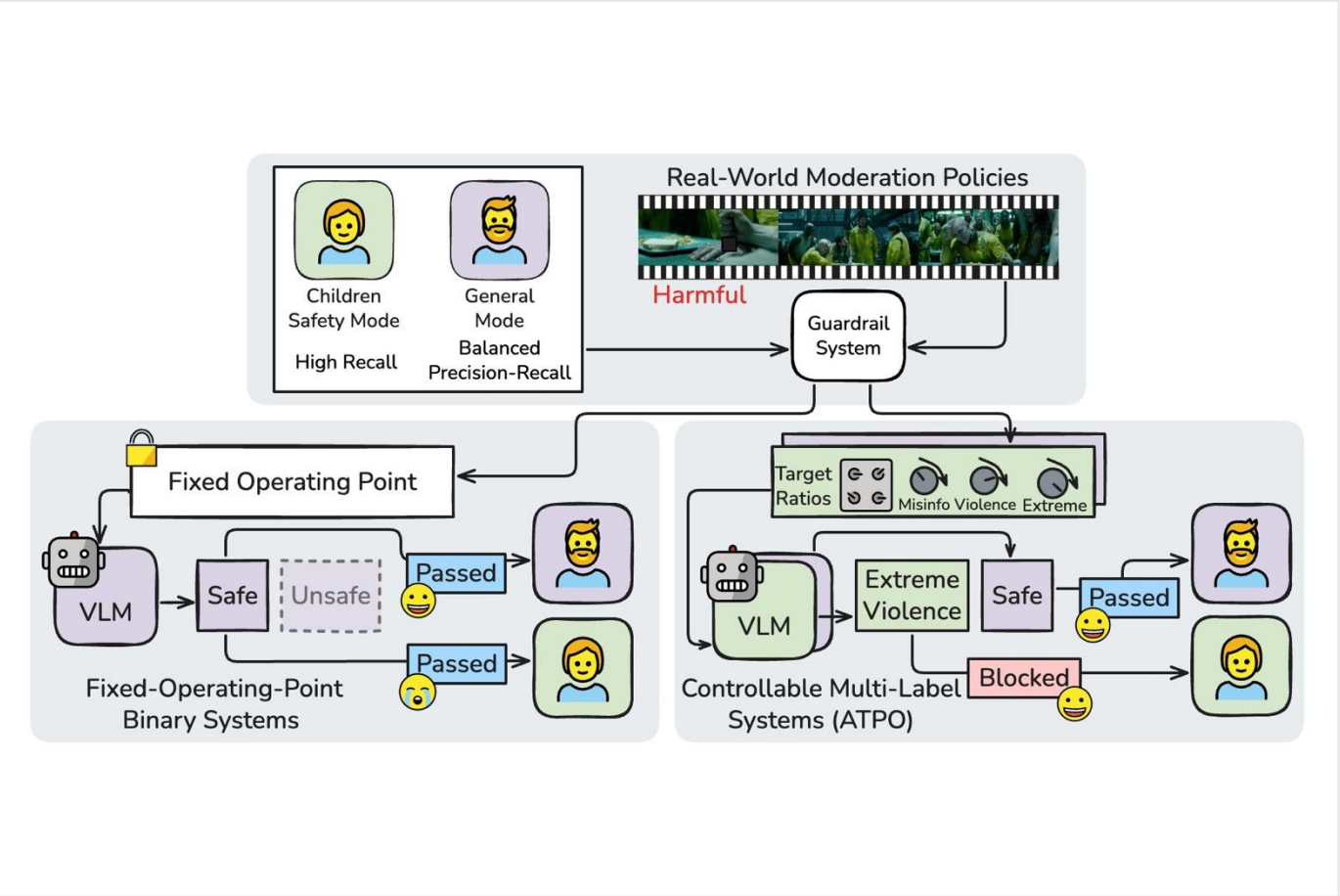
ATPO: from fixed objectives to controllable multi-label detection

Limitations of existing methods

- Binary labels cannot distinguish different risk categories.
- Fixed training objectives cannot adapt to the desired operating point.

ATPO's improvements

- Multi-label outputs identify the relevant risk categories.
- Training objectives adapt to global or category-specific operating points.



Adaptive Tversky Reward weights FP and FN

$$R_t = \frac{TP_t}{TP_t + \alpha_t FP_t + \beta_t FN_t}$$

Penalty for False Positives

Penalty for False Negatives

Dynamically adjusted during ATPO training

α_t penalty on false positives

A larger α_t penalizes false positives more strongly.

β_t penalty on false negatives

A larger β_t penalizes missed harmful labels more strongly.

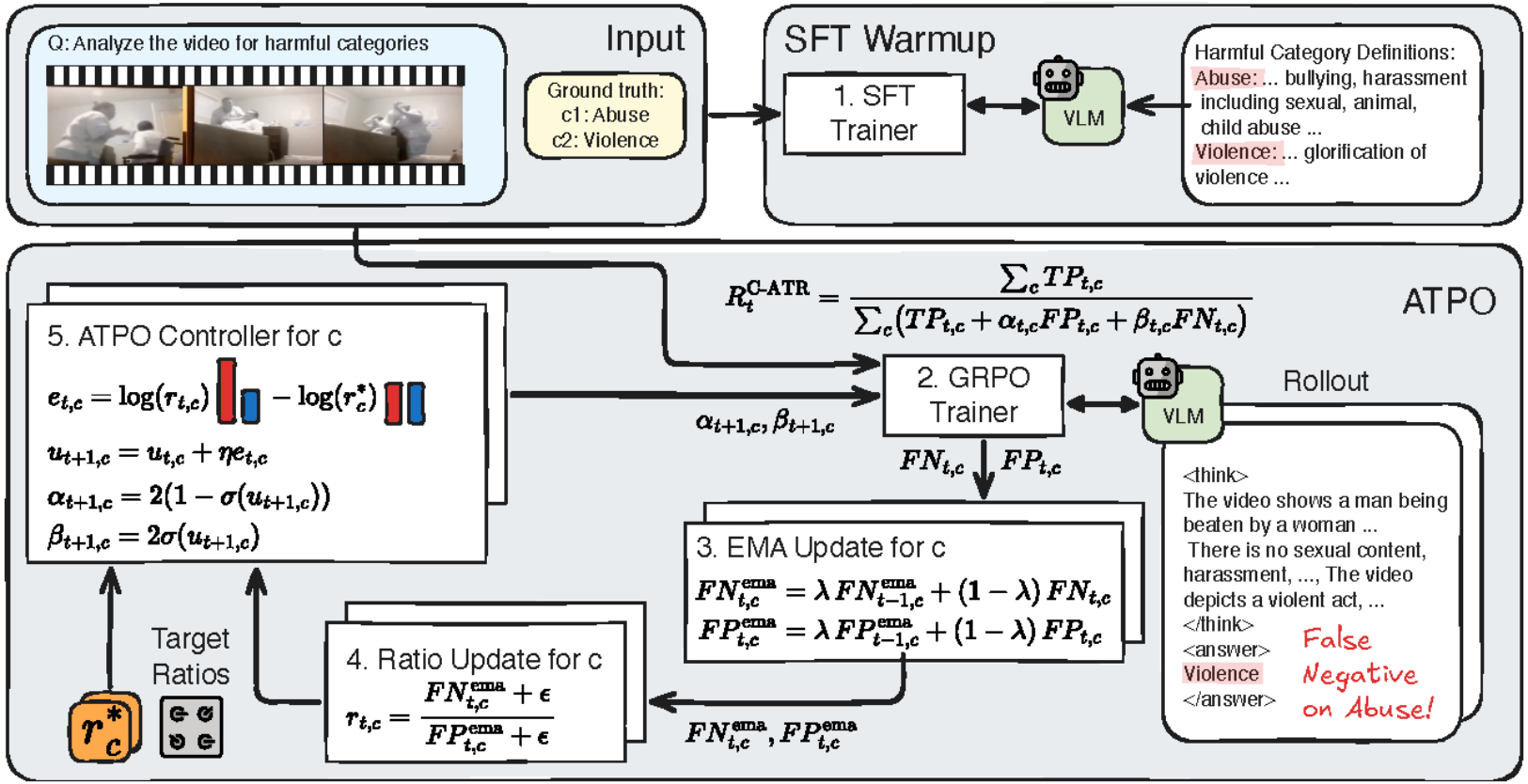
The controller adjusts α_t and β_t during training dynamically, with $\alpha_t + \beta_t = 2$.

A feedback controller updates the reward

1. **SFT warmup**
   Initialize multi-label behavior

feedback loop

2. **GRPO rollouts**
   Optimize with adaptive reward
3. **EMA update**
   Track category-wise FNs and FPs
4. **Ratio update**
   Estimate FN / FP from smoothed error counts
5. **ATPO controller**
   Adjust α and β using the target error ratio



ATPO improves results on both video benchmarks

SafeWatch-Bench-Real

Model	Jaccard ↑	Binary F1 ↑
Qwen2.5-VL-7B	40.66	65.94
Qwen3-VL-235B-A22B	55.85	80.29
InternVL3.5-38B	60.61	85.80
VAD-R1*	46.36	81.13
SafeWatch-8B†	<u>70.69</u>	<u>89.04</u>
ATPO-G-7B	**75.44**	**94.82**

XD-Violence

Model	Jaccard ↑	Binary F1 ↑
Qwen2.5-VL-7B	76.54	90.21
Qwen3-VL-235B-A22B	82.53	<u>95.36</u>
InternVL3.5-38B	<u>83.38</u>	93.59
VAD-R1*	74.12	89.69
SafeWatch-8B†	79.84	86.80
ATPO-G-7B	**88.17**	**96.69**

+34.78 Jaccard points over the backbone on SafeWatch-Real

Bold: best; underlined: second best. † SafeWatch uses full videos and clips. * VAD-R1 has additional benchmark training.

Adaptive rewards improve detection and transfer

Reward attribution

SafeWatch-Real

Method	Jaccard ↑	Binary F1 ↑
SFT (1 epoch)	44.33	63.53
GRPO (exact match)	63.36	92.30
GRPO (static Tversky)	<u>70.09</u>	<u>93.97</u>
ATPO (G-ATR)	**75.44**	**94.82**

+5.35 Jaccard over static reward

Distribution shift

VHD11K-Video-Real, without target fine-tuning

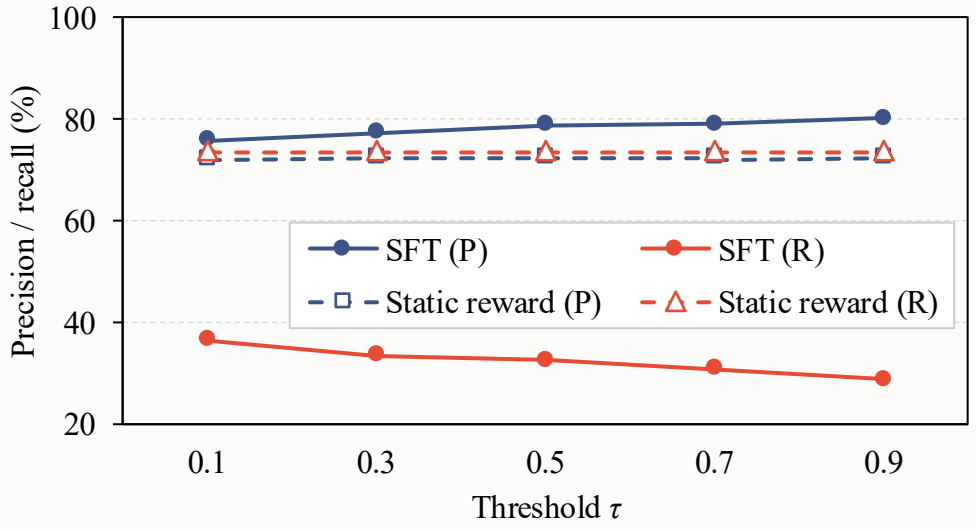
Method	Jaccard ↑	Binary F1 ↑
SFT (1 epoch)	<u>63.64</u>	<u>80.40</u>
SFT (3 epochs)	57.63	76.00
GRPO (static Tversky)	59.03	78.20
ATPO (C-ATR)	**68.50**	**84.00**

ATPO-C leads the reported VHD11K baselines.

Training targets reach additional operating points

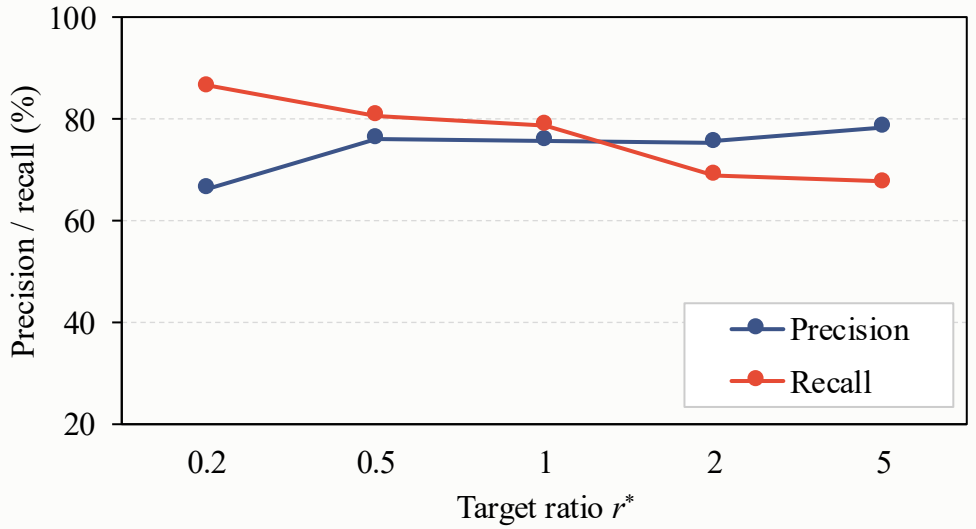
SafeWatch-Bench-Real · Micro Precision / Recall

Fixed models: threshold sweep



28.61–36.36 SFT recall range

ATPO: separate training runs



67.62–86.55 ATPO recall range

Comparison with the reported global token-probability threshold sweep (Appendix K).

C-ATR cuts each category's targeted error by 34% and 74%

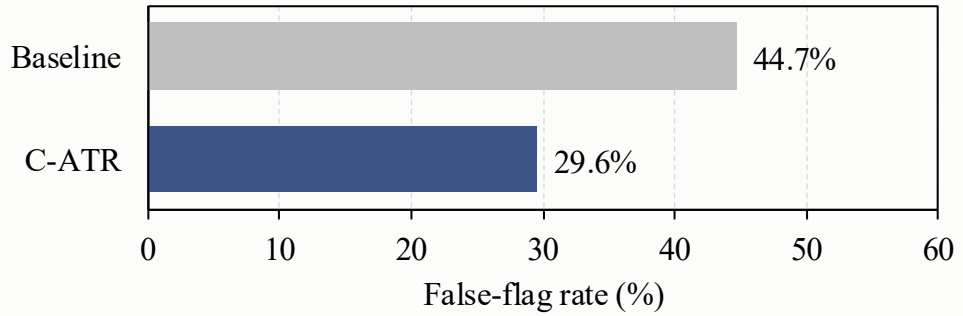
SafeWatch-Real (%) · Baseline: all targets = 1 · Changed: C4 = 5, C6 = 0.2

C4: Misinformation

Target 5.0

Precision priority

−34% false-flag rate vs. baseline



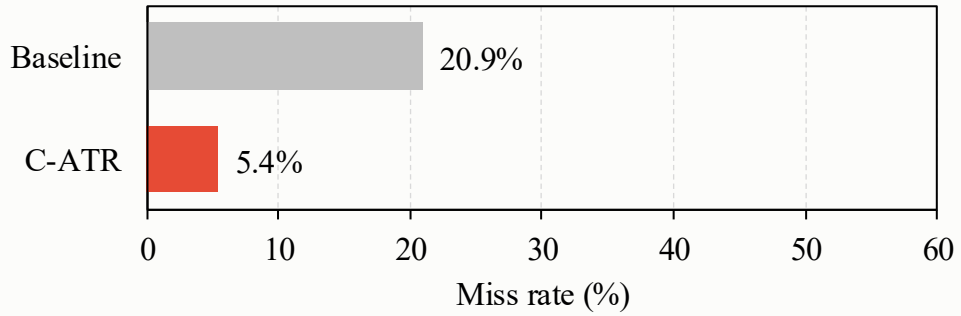
Recall holds: 66.67 → 67.65

C6: Extremism

Target 0.2

Recall priority

−74% miss rate vs. baseline



Trade-off: precision falls 80.31 → 31.77

False-flag rate = 100 − precision; miss rate = 100 − recall. Reductions are relative to the baseline rate.

ATPO makes the error trade-off part of training.

01 Multi-label predictions retain category information.

02 Adaptive rewards improve on matched static-reward GRPO.

03 Global and category-specific targets steer precision and recall.

Paper and code: bruceyg.github.io/ATPO-project-page/

Thank you