SenBen: Sensitive Scene Graphs for Explainable Content Moderation

Fatih Cagatay Akyon[12] · Alptekin Temizel[1]

https://senben.kim

CVPR JUNE 3-7, 2026 DENVER COLORADO

10th ABAW Workshop · CVPR 2026

Motivation & Contributions

- Today's content moderation outputs flat labels.
- Existing benchmarks (LSPD, NudeNet) are saturated at 0.95+ F1.
- Our idea: a scene graph IS the explanation.

male:aggression →holding→ knife ⇒ violence

- Three contributions:
  - SenBen · 13,999 frames · 157 movies · 16 tags.
  - Recipe: Suffix + VAR + Q2L → +6.4 pp over CE.
  - 241 M student beats all local VLMs and safety APIs.

SenBen Dataset

- 13,999 frames · 157 movies
- 50/50 sensitive vs general
- 25 obj · 28 attr · 14 pred
- 5 categories · 16 tags

Split	Frames	Movies
Train	9,999	95
Val	2,000	31
Test	2,000	31
Total	**13,999**	**157**

Tag Distribution · 16 tags

Category	Tag	Total
Violence	violence	2,521
	gore	438
Substances	drugs legal	1,402
	drugs illegal	330
Immodesty	immodesty	1,066
	nudity	520
	nudity implied	500
	nudity art	199
Sexual	sex. suggestive	794
	kissing	344
	sexual activity	289
	sex implied	233
Other	medical graphic	248
	medical proc.	237
	bodily functions	131
	vulgar gestures	78

Method: Multi-Task Distillation

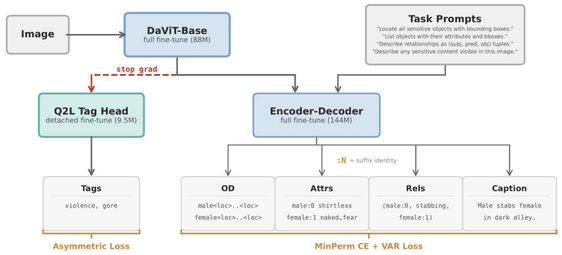


Fig. 1. DaViT vision encoder → Q2L tag head + encoder-decoder for 4 SG tasks.

- Florence-2 + Q2L head = **241 M params**.
- Five tasks share one encoder.
- Beam B=3, **733 ms/frame, 1.2 GB VRAM**.

Key Ideas

(a) Suffix-based object identity

- Replace bbox tokens with **:N suffixes**.
- Shorter sequences (**17–27 %**), cross-task consistency.
- Biggest single ablation win: **+3.8 pp SenBen-Recall**.

Attrs : male:**0** shirtless female:**1** naked, fear
Rels : (male:**0**, stabbing, female:**1**)

(b) VAR Loss · Vocabulary-Aware Recall

$$L = L\_CE + \lambda \cdot (1 - R\_s)^{\gamma}$$

- Penalty only at sensitive-vocab positions.
- λ = 0.1, γ = 2, 200-step warmup.
- +3.4 pp SenBen-Recall.

Decoupled Q2L Tag Head

- Tag grads **orthogonal** to scene-graph grads.
- Q2L head: 16 label queries, stop-gradient, **Asymmetric Loss**.
- Two configs: balanced (γ^-=4, γ^+=1) and aggressive (γ^-=7, γ^+=0).
- Effect: **+7.8 pp tag F1**. No scene-graph regression.

Results: Latency vs Accuracy

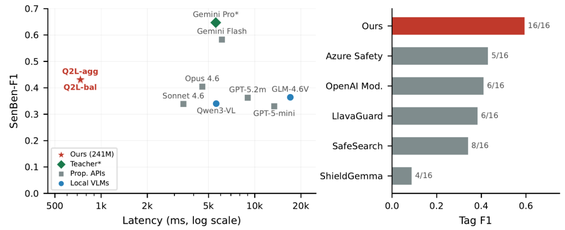


Fig. 2. SenBen-F1 vs latency on 2k frames.

Best speed × accuracy × coverage trade-off

Ablation · Each Ingredient Counts

System	SB-R	SB-F1	Tag F1
CE (baseline)	.349	.389	.544
+ Suffix	.366	.406	.509
+ VAR	.376	.408	.509
+ MinPermCE	.386	.415	.532
+ Label smoothing	.385	.419	.533
+ Scheduled sampling	.392	.422	.516
+ Q2L bal. (full)	**.413**	**.428**	**.594**

- **Drop Suffix → −3.8 pp · Drop VAR → −3.4 pp.**

Comparison vs VLMs

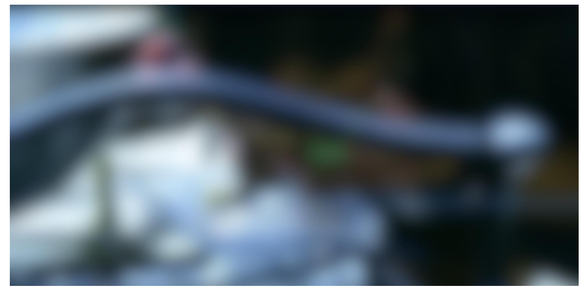
Model	Params	Tag F1	R(obj)	SB-F1
Gemini 3 Pro*	—	.806	.295	.647
Gemini 3 Flash	—	.784	.271	.583
Q2L-bal (ours)	**241M**	**.594**	**.420**	**.428**
Q2L-agg (ours)	**241M**	**.457**	**.409**	**.431**
Claude Opus 4.6	—	.658	.082	.404
GLM-4.6V (reas.)	10.3B	.492	.123	.364
GPT-5.2 (med. r.)	—	.608	.072	.362
Qwen3-VL-8B	8.3B	.469	.104	.340

** Teacher. Generated initial labels, human-corrected.*

vs Commercial Safety APIs

Model	Tags	Tag F1	Safe F1
Q2L-bal (ours)	**16/16**	**.594**	**.847**
Q2L-agg (ours)	**16/16**	**.457**	**.835**
Azure Cont. Safety	5/16	.430	.504
OpenAI Mod.	6/16	.411	.664
LlavaGuard 1.2	6/16	.384	.583
Google SafeSearch	8/16	.341	.476
ShieldGemma 2	4/16	.089	.161

Qualitative · Scene Graphs



male →snorting→ powder. Tag: **drugs_illegal**.

Conclusions & Takeaways

- **Scene graphs > flat labels.**
  First grounded benchmark for explainable moderation.
- **Distillation pays off.**
  241M beats every local VLM and every safety API.
- **Two ingredients carry the recipe.**
  Suffix (+3.8 pp) and VAR (+3.4 pp).
- **Decoupling tag head matters.**
  +7.8 pp tag F1, no SG regression.
- **ABAW relevance.**
  Affective attributes grounded to actors.

The output is the explanation.

Affiliations
[1] Graduate School of Informatics, METU, Turkiye
[2] Ultralytics, Inc., USA
10th ABAW Workshop · CVPR 2026

Paper & Dataset
Project: https://senben.kim
Dataset: https://huggingface.co/datasets/fcakyon/senben

Dataset · F.C.Akyon · A.Temizel

The output is the explanation.