SafeSocial: On-device detection of social-comparison triggers in Instagram-style images

June 17, 2026

Authors: Emre Sokullu (WebBrain, emre@webbrain.one)

Manuscript status: Draft; not yet supported by human-ground-truth evaluation

Summary

This draft reports a teacher-distillation study for local detection of social-comparison triggers in Instagram-style images. The strongest research baseline is SigLIP2, while the strongest compact deployment candidate is a 20-epoch EfficientNet-Lite0 model trained on core Instagram-style data, a partial Kaggle expansion, and a capped 20% COCO auxiliary mix. All performance numbers measure agreement with the Qwen teacher rather than human validity.

Abstract

Social-comparison content on image-centric platforms may contribute to jealousy, fear of missing out, status anxiety, and body-comparison pressure. A practical protective system should run locally, because uploading a user's feed images to a server would create privacy and platform-trust concerns. We study whether compact image classifiers can approximate a large vision-language teacher for detecting social-comparison triggers in Instagram-style images. A Qwen qwen3.6-35b-a3b teacher labels public Instagram-style image-caption data using a reduced, visually grounded taxonomy covering romance jealousy, social FOMO, luxury/status, travel/lifestyle, body/beauty comparison, achievement/status, social proof/popularity, exclusivity/access, none, and uncertain. We compare frozen vision encoders for research quality and compact trainable backbones for browser/mobile deployment. SigLIP2 reaches 0.7660 micro F1, 0.6236 macro F1, and 0.8977 label accuracy against teacher labels on the 20k Instagram-style test split. The strongest compact candidate is EfficientNet-Lite0 trained for 20 epochs on core Instagram/HF data plus a partial Kaggle expansion and a capped 20% Karpathy COCO auxiliary mix. With tuned per-label thresholds it reaches 0.6817 micro F1, 0.5126 macro F1, 0.2952 exact match, and 0.8684 label accuracy on the primary merged test split. On a separate 500-image COCO out-of-domain holdout, auxiliary training improves label accuracy from 0.8246 to 0.9176 and exact match from 0.2360 to 0.5480 versus the pre-aux compact checkpoint. These results support a plausible path to privacy-preserving browser and mobile filtering, but they remain teacher-agreement results and require human-verified evaluation before product or clinical claims.

Introduction

Image feeds can concentrate social-comparison cues: romantic display, luxury consumption, travel lifestyle, body presentation, achievement status, popularity signals, and exclusive access. Existing interventions often rely on user self-control, platform-side ranking, or server-side moderation. The product goal here is different: a local classifier that detects high-risk visual triggers on the user's device and can blur, hide, dim, or annotate content without sending images to a remote classifier.

The core research question is whether a compact on-device image model can approximate a large vision-language teacher well enough to support a privacy-preserving browser extension or mobile wrapper. The study is intentionally framed as teacher distillation. The current benchmark does not estimate psychological harm, user outcomes, or human-judged trigger validity. Instead, it measures whether smaller students can reproduce a consistent teacher taxonomy over Instagram-style images.

This framing matters for a potential high-impact submission. The technical result is strongest as an on-device machine-learning and human-computer-interaction prototype. A stronger behavioral or mental-health claim will require a human-labeled evaluation set and later user studies.

Results

A first broad taxonomy included envy and financial FOMO. Those labels were removed because they were visually underdetermined or rare in the sampled Instagram-style images. The reduced core taxonomy improved the tractability of the task and focused the classifier on visually grounded triggers.

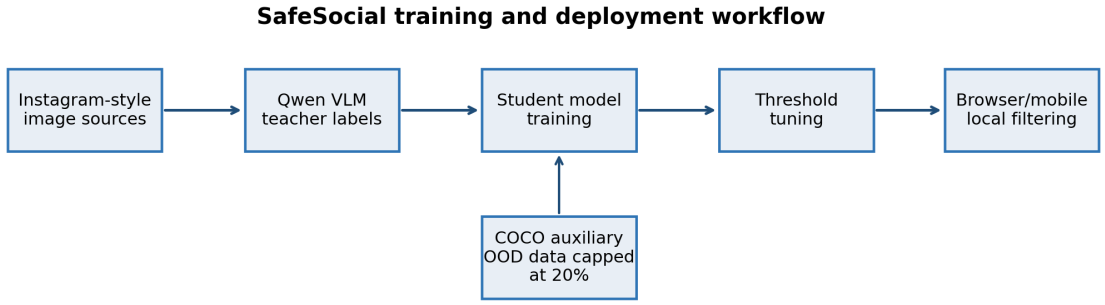

Figure 1. SafeSocial training and deployment workflow. Primary Instagram-style data remains the benchmark domain, while COCO/Flickr-style data is capped and evaluated separately as auxiliary OOD evidence.

Table 1. Research backbone comparison on the 20k core teacher-labeled test split.

Backbone	Micro F1	Macro F1	Exact	Label acc.
SigLIP2	0.7660	0.6236	0.3747	0.8977
MetaCLIP2	0.7354	0.5933	0.3192	0.8817
DINOv2	0.6785	0.5374	0.2431	0.8509
Small CNN	0.4050	0.3172	0.0240	0.6396

SigLIP2 is the strongest research baseline, reaching 0.7660 micro F1 and 0.6236 macro F1. MetaCLIP2 and DINOv2 remain strong but trail SigLIP2 on both micro and macro F1. The compact CNN baseline is not adequate for this task.

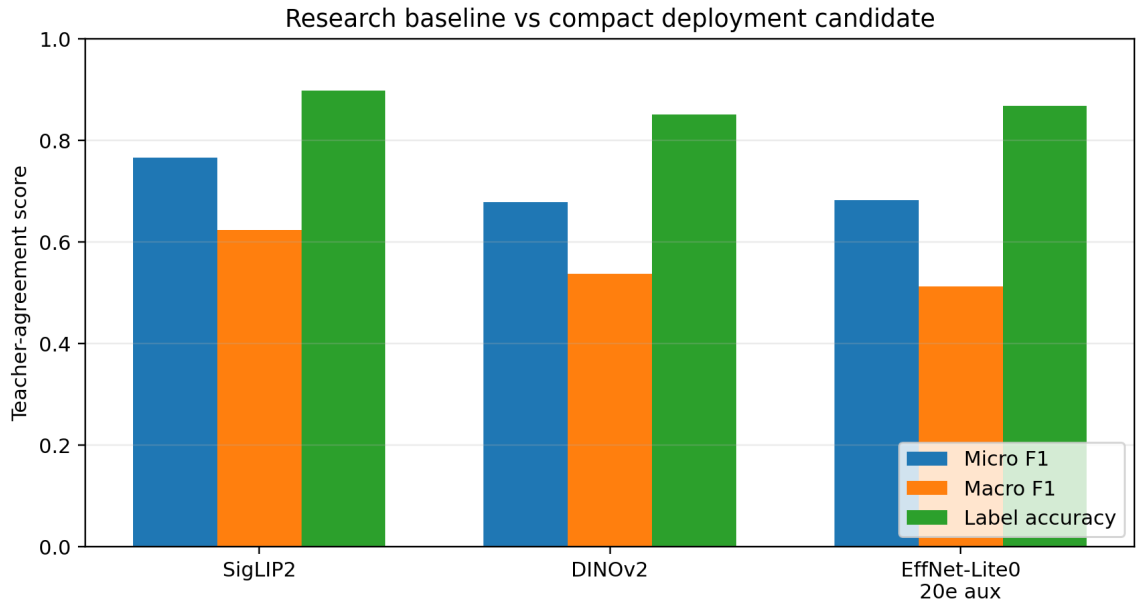

Figure 2. Teacher-agreement performance for a research baseline, a self-supervised baseline, and the current compact deployment candidate.

Table 2. Compact deployment progression on the primary merged Instagram-domain test split.

Model / setting	Micro F1	Macro F1	Exact	Label acc.
MobileNetV3-Small / Core 20k	0.5869	0.4546	0.1426	0.7930
EfficientNet-Lite0 / Core 20k	0.6027	0.4615	0.1726	0.8077
EfficientNet-Lite0 / Core + Kaggle, tuned	0.6710	0.5019	0.2793	0.8576
EfficientNet-Lite0 / Core + Kaggle + 20% COCO aux, 10 epochs, tuned	0.6703	0.5034	0.2903	0.8609
EfficientNet-Lite0 / Core + Kaggle + 20% COCO aux, 20 epochs, tuned	0.6817	0.5126	0.2952	0.8684

The compact path improves in three stages. First, EfficientNet-Lite0 is stronger than MobileNetV3-Small on the same 20k core split. Second, adding the partial Kaggle Instagram archive and tuning per-label thresholds substantially improves decision quality. Third, capped auxiliary COCO training and a longer 20-epoch schedule further improve the deployment candidate without changing the primary validation or test domain. The current preferred compact checkpoint reaches 0.6817 micro F1, 0.5126 macro F1, and 0.8684 label accuracy.

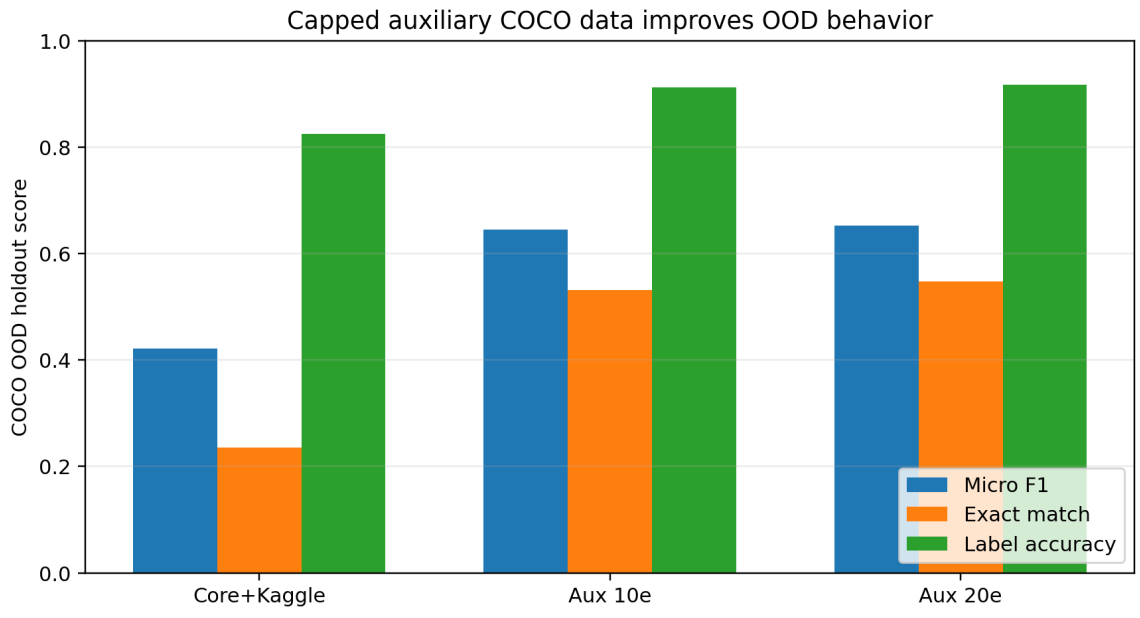

Figure 3. The capped 20% COCO auxiliary mix improves behavior on the separate 500-image COCO OOD holdout.

Table 3. COCO OOD robustness on a 500-image holdout, evaluated separately from the Instagram-domain benchmark.

Model / setting	Micro F1	Macro F1	Exact	Label acc.
Core + Kaggle, tuned	0.4211	0.2654	0.2360	0.8246
Core + Kaggle + 20% COCO aux, 10 epochs, tuned	0.6456	0.3252	0.5320	0.9124
Core + Kaggle + 20% COCO aux, 20 epochs, tuned	0.6526	0.3466	0.5480	0.9176

The pre-aux compact model produces too many false positives on ordinary general images. With the capped COCO auxiliary mix, OOD exact match improves from 0.2360 to 0.5480 and label accuracy improves from 0.8246 to 0.9176. This is the strongest evidence that auxiliary OOD data is useful, but it should not be mixed into the primary Instagram benchmark.

Table 4. Label distribution in the 5,000-row Karpathy COCO auxiliary run.

Label	Count	Share
none	3368	67.4%
travel_lifestyle	861	17.2%
achievement_status	347	6.9%
luxury_status	305	6.1%
body_beauty_comparison	264	5.3%
social_proof_popularity	253	5.1%
social_fomo	155	3.1%
exclusivity_access	65	1.3%
romance_jealousy	47	0.9%

Methods

The primary data source is kkcosmos/instagram-images-with-captions, from which 20,000 images were exported and 19,991 valid teacher-labeled rows were retained. The Kaggle Instagram archive prithvijaunjale/instagram-images-with-captions contributed 2,641 additional valid teacher-labeled rows from a partial expansion. The merged split contains 18,106 training rows, 2,263 validation rows, and 2,263 test rows. The Karpathy Deep Image Sentences COCO subset contributed 5,000 teacher-labeled OOD rows; 4,500 were used as capped auxiliary training data and 500 were held out for separate OOD evaluation.

Teacher labels were generated by qwen3.6-35b-a3b served through a local OpenAI-compatible vLLM endpoint. The teacher returned multi-label JSON and rationales under a rubric that prioritizes visible evidence. The reduced taxonomy is romance_jealousy, social_fomo, luxury_status, travel_lifestyle, body_beauty_comparison, achievement_status, social_proof_popularity, exclusivity_access, none, and uncertain.

Research students freeze pretrained image encoders and train a multi-label head. Compact deployment candidates train mobile-friendly backbones end to end. The strongest deployment checkpoint uses timm tf_efficientnet_lite0, 20 training epochs, batch size 64, and per-label thresholds tuned on the validation split. Tuned thresholds are 0.75 for romance_jealousy, 0.80 for social_fomo, 0.45 for luxury_status, 0.70 for travel_lifestyle, 0.35 for body_beauty_comparison, 0.90 for achievement_status, 0.60 for social_proof_popularity, 0.60 for exclusivity_access, 0.75 for none, and 1.00 for uncertain.

Metrics are micro F1, macro F1, exact match, hamming loss, and label accuracy, where label accuracy is 1 minus hamming loss. Macro F1 is emphasized because rare triggers such as romance jealousy or exclusivity access can be hidden by common classes and negative examples. All metrics in this manuscript are teacher-agreement metrics unless explicitly marked otherwise.

Prototype implementation

The browser prototype is a Chrome/Edge Manifest V3 extension. It observes Instagram media elements, loads a local ONNX bundle when present, falls back to deterministic local mock scoring when the runtime is unavailable, and applies blur, hide, dim, or warning overlays. The threshold control acts as a sensitivity offset around tuned per-label thresholds. Images remain local to the browser.

The mobile prototype is an Expo application with a WebView pinned to Instagram domains. The injected DOM observer sends asynchronous scoring requests to React Native through postMessage. The React

Native adapter currently returns deterministic scores but is structured as the integration point for ONNX Runtime React Native, TensorFlow Lite, or TensorFlow.js React Native.

Discussion

The main scientific result is that a compact EfficientNet-Lite0 student can capture a meaningful fraction of the teacher's social-trigger taxonomy while remaining small enough for browser and mobile deployment experiments. The main engineering result is that the same exported model metadata can drive both a browser extension and a mobile wrapper architecture. The main methodological result is that auxiliary OOD images should be capped and reported separately rather than used to inflate primary benchmark numbers.

The results also show a clear quality gap. SigLIP2 remains stronger than the compact deployment model on macro F1, and exact match is still below 0.30 on the primary merged test set. Some labels are difficult: romance_jealousy and exclusivity_access have lower F1 than body_beauty_comparison or luxury_status. This is expected because some triggers depend on context, captions, or user history rather than the visual content alone.

Limitations

- The evaluation set is teacher-labeled. It measures student-teacher agreement, not human validity.
- The taxonomy is subjective and requires human rater validation before product or health claims.
- The public Instagram-style data sources have unclear licensing histories and should be replaced or supplemented with opt-in or clearly licensed production data.
- The Kaggle expansion is partial and likely skewed toward celebrity and beauty content.
- COCO/Flickr-style data is out-of-domain for Instagram and should be used only as capped auxiliary robustness data.
- The mobile runtime bridge is scaffolded but native model execution is not yet implemented.

Data availability

The experimental pipeline writes manifests, teacher labels, splits, checkpoints, predictions, and metrics under image_classifier/runs in the SafeSocial repository. Public source datasets are referenced below. Because the source data may carry licensing and platform-policy constraints, redistributed image files are not included in the manuscript artifacts.

Code availability

The repository contains the teacher-labeling pipeline, multi-label training scripts, threshold tuning, ONNX export, browser extension prototype, Expo mobile prototype, and paper notes. The current branch is codex/social-trigger-pipeline.

Ethics and privacy

The intended product direction is privacy-preserving local filtering. Images should not leave the user's browser or phone during inference. The classifier should be presented as a user-controlled filtering aid, not as a clinical diagnostic or platform moderation authority. Before deployment, the taxonomy and thresholds should be reviewed with human raters and evaluated for disparate false-positive behavior across identity, body type, culture, and content genre.

Competing interests

The authors declare no competing interests, pending author review.

References

1. kkcosmos. instagram-images-with-captions dataset card. Hugging Face. https://huggingface.co/datasets/kkcosmos/instagram-images-with-captions

2. Jaunjale, P. Instagram Images with Captions. Kaggle. https://www.kaggle.com/datasets/prithvijaunjale/instagram-images-with-captions

3. Karpathy, A. Deep Image Sentences: Flickr8K, Flickr30K and MSCOCO caption data. https://cs.stanford.edu/people/karpathy/deepimagesent/

4. Google. SigLIP2 model card: google/siglip2-base-patch16-224. Hugging Face. https://huggingface.co/google/siglip2-base-patch16-224

5. Meta AI. MetaCLIP 2 model card: facebook/metaclip-2-worldwide-s16. Hugging Face. https://huggingface.co/facebook/metaclip-2-worldwide-s16

6. Meta AI. DINOv2 model card: facebook/dinov2-small. Hugging Face. https://huggingface.co/facebook/dinov2-small

7. Howard, A. et al. Searching for MobileNetV3. Proceedings of ICCV (2019).

8. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of ICML (2019).

9. Wightman, R. PyTorch Image Models (timm). https://github.com/huggingface/pytorch-image-models

10. ONNX Runtime Web documentation. Microsoft. https://onnxruntime.ai/docs/tutorials/web/

11. Nature. Initial submission. https://www.nature.com/nature/for-authors/initial-submission

12. Nature. Formatting guide. https://www.nature.com/nature/for-authors/formatting-guide