Layer-Wise CLIP Feature Fusion for Synthetic Image Detection

Jeng Wen Joshua Lean[1]

[1]National Tsing Hua University, Hsinchu, Taiwan

Abstract

For MediaEval 2026 Synthetic Image Detection Task A, we propose a system that uses OpenCLIP ViT-L/14 as a visual encoder and learns a compact classifier over layer-wise class-token features, allowing intermediate representations to contribute to real-versus-synthetic decisions. We compare a constrained run trained on COCO real images and Corvi synthetic images with an open run that additionally uses TrueFake social images. On local validation data, the open run improves F1 from 0.7555 to 0.7916 and ROC AUC from 0.8114 to 0.8643. On the official challenge evaluation, it improves F1 from 0.7013 to 0.7150 and ROC AUC from 0.7589 to 0.7686. These results suggest that intermediate CLIP features are useful for synthetic image detection, while broader social-image training data mainly improves recall.

1. Introduction

The MediaEval 2026 Synthetic Images task asks participants to detect whether an image is real or synthetic under realistic online conditions, including compression, resizing, and cropping [1]. Task A requires at least one constrained run and one open run, with analysis of how training data or model choices change behavior. This paper reports our standard Working Notes submission for Task A.

Our system is motivated by two observations. First, CLIP-style encoders provide strong transferable visual features for generated-image detection [2]. Second, the final embedding of a vision transformer is not necessarily the only useful representation for forensics: intermediate blocks can retain traces that are attenuated by the final semantic representation [3]. We therefore train a compact fusion head over class-token features from all OpenCLIP ViT-L/14 visual transformer blocks.

The main question we analyze is how this layer-wise detector behaves when the training data is broadened beyond the constrained subset. The open-data run adds TrueFake social images [4], which are closer in style to online media than the COCO/Corvi-only constrained run. We report local validation, official evaluation, and robustness results to show both the benefit and the remaining failure modes.

2. Method

Given an RGB image x, the model predicts whether x is real or synthetic. Images are converted to RGB, normalized with CLIP image statistics, and cropped to 224×224. Training uses random crops, horizontal flips, color jitter, mild rotation, JPEG recompression, and downsample-upsample degradation. These augmentations target the transformation sensitivity highlighted by recent synthetic-image detection work [5, 6].

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
joshualeanjw@gmail.com (J. W. J. Lean)

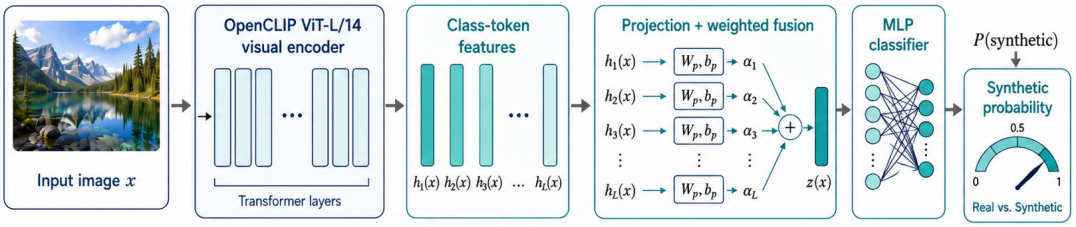

Figure 1: Method overview: images are encoded by CLIP, intermediate class-token features are fused with learned block weights, and a binary classifier outputs a synthetic-image probability.

The detector uses a pretrained OpenCLIP ViT-L/14 visual encoder [7]. Let $h_l(x)$ be the class-token feature from transformer block l. Each block feature is projected to a shared space, and the model learns a softmax-normalized scalar weight α_l for each block:

$$z(x) = \sum_{l=1}^{L} \alpha_l (W_p h_l(x) + b_p).$$

A small multilayer classifier maps $z(x)$ to a synthetic-image logit. The model is trained with binary cross-entropy. In the reported runs, only the classifier head and the final two CLIP visual blocks are fine-tuned; the rest of the visual encoder remains frozen. Validation probabilities are used to choose the F1-maximizing global threshold for each run.

3. Experimental Setup

We evaluate two training regimes. The constrained run uses 75,000 COCO train2017 real images [8] and 75,000 Corvi latent-diffusion synthetic images [9]. The task's constrained data also includes Wang et al. GAN images [10]; our submitted constrained configuration uses the COCO/Corvi subset. The open run adds 60,000 TrueFake social real images and 60,000 TrueFake social synthetic images [4], for 135,000 real and 135,000 synthetic training images.

Both runs use the provided local validation split of 5,000 real and 5,000 synthetic in-the-wild images. Official evaluation is performed on the hidden 10,000-image challenge test set. We report accuracy, precision, recall, F1, ROC AUC, and average precision (AP). F1 and the confusion matrices use the calibrated global threshold, while ROC AUC and AP are threshold-independent.

Both runs use OpenCLIP ViT-L/14 with OpenAI pretrained weights, batch size 16, three training epochs, AdamW with weight decay 10^{-4}, a head learning rate of 3×10^{-4}, and a backbone learning rate of 10^{-5} for the unfrozen blocks. The best checkpoint is selected by validation ROC AUC, with AP as a tie-breaker. The random seed is 1337.

4. Results

Table 1 summarizes local validation and official challenge performance. On local validation, the open run improves every reported ranking and thresholded metric. The largest thresholded gain is recall, which increases from 0.8288 to 0.8964.

The official challenge scores are lower than local validation scores, indicating a distribution shift between the development split and the hidden evaluation data. The open run remains the best run by official F1 and ROC AUC, improving F1 by 0.0137 and ROC AUC by 0.0097 over the constrained run. The official gain is again recall-driven: recall rises from 0.7332 to 0.7612,

Table 1
Main results for constrained and open-data runs. AP denotes average precision.

Split	Run	Acc.	Prec.	Rec.	F1	AUC	AP
Local	Constr.	0.7318	0.6941	0.8288	0.7555	0.8114	0.7963
Local	Open	0.7640	0.7087	0.8964	0.7916	0.8643	0.8541
Official	Constr.	0.6877	0.6720	0.7332	0.7013	0.7589	0.7665
Official	Open	0.6966	0.6741	0.7612	0.7150	0.7686	0.7607

Table 2
Selected local robustness results.

Run	JPEG F1	JPEG AUC	Launder F1	Launder AUC	Small F1
Constr.	0.7247	0.7716	0.7162	0.7826	0.5833
Open	0.7631	0.8385	0.7379	0.8212	0.7156

while precision changes only from 0.6720 to 0.6741. AP is slightly lower for the open run, so the additional social data does not uniformly improve all ranking behavior.

Table 2 shows selected robustness checks: JPEG quality 85, a resize/crop laundering pipeline, and images whose shorter side is below 512 pixels. The open run improves all three slices, most notably the small-image bucket. However, performance remains lower than clean validation performance, which reinforces the task premise that online transformations are a central challenge rather than a secondary nuisance.

5. Analysis and Conclusion

Layer-wise CLIP fusion provides a practical representation for synthetic-image detection without full end-to-end fine-tuning of the large visual encoder. The results are consistent with earlier CLIP-based detection work [2] and with the hypothesis that intermediate encoder blocks retain useful forensic information [3].

The constrained-to-open comparison shows that broader social-image training data helps most through recall. On the official evaluation, the open run reduces synthetic false negatives from 1,334 to 1,194, but increases real-image false positives from 1,789 to 1,840. This is useful for a recall-oriented task metric, but it also means deployment would need threshold selection that accounts for false-positive cost. The very small calibrated thresholds further indicate that raw probabilities should not be treated as well-calibrated probabilities across new target distributions.

Overall, our submission favors a lightweight and reproducible route: expose intermediate CLIP representations, train a compact fusion head with transformation-aware augmentation, and calibrate the decision threshold on the target evaluation protocol. Future work should evaluate the same model under more platform-specific transformations and analyze false positives by visual content and acquisition source.

Declaration on Generative AI

During the preparation of this work, the author used OpenAI ChatGPT/Codex for revision planning, grammar and spelling checks, paraphrasing and rewording, and LaTeX formatting

assistance. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] D. Karageorgiou, O. Papadopoulou, S. Papadopoulos, C. Koutlis, H. Mareen, E. Gavves, Synthetic images: Advancing detection and localization of generative ai used in real-world online images, in: Proc. of the MediaEval 2026 Workshop, Amsterdam, Netherlands and Online, 2026. Task overview paper, forthcoming.

[2] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, L. Verdoliva, Raising the bar of ai-generated image detection with clip, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, pp. 4356–4366. doi:10.1109/CVPRW63382.2024.00439.

[3] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, in: Computer Vision – ECCV 2024, Springer, 2024, pp. 394–411. doi:10.48550/arXiv.2402.19091.

[4] S. Dell'Anna, A. Montibeller, G. Boato, Truefake: A real world case dataset of last generation fake images also shared on social networks, arXiv preprint arXiv:2504.20658, 2025. doi:10.48550/arXiv.2504.20658.

[5] P. Grommelt, L. Weiss, F.-J. Pfreundt, J. Keuper, Fake or jpeg? revealing common biases in generated image detection datasets, in: Computer Vision – ECCV 2024 Workshops, Springer, 2025, pp. 80–95. doi:10.1007/978-3-031-92089-9_6.

[6] O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, F. Feng, Improving synthetic image detection towards generalization: An image transformation perspective, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025. doi:10.1145/3690624.3709392.

[7] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, Openclip, 2021. doi:10.5281/zenodo.5143773.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision – ECCV 2014, Springer, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[9] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, in: ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing, 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10095167.

[10] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8695–8704.