A Progressive Pipeline from CNNs to CLIP-Based Feature Extraction for Synthetic Image Detection at MediaEval 2026

Maham Junaid[1], Maryam Ihsan[1], Hamna Usman[1] and Atif Tahir[1]

[1]*Institute of Business Administration (IBA) Karachi, Pakistan*

Abstract

This paper describes our constrained and open-run submissions to Subtask A of the MediaEval 2026 Synthetic Images Challenge [1], which requires binary classification of social media images as real or AI-generated. For the constrained run, we trained exclusively on the provided ITW-SM validation set (9,999 images), progressing through a multi-phase pipeline of CNN baselines (EfficientNet-B4, ConvNeXt-Tiny, a dual-stream variant) before settling on CLIP ViT-L/14 with a lightweight probe. For the open run, we added the CIFAKE dataset and evaluated two CLIP backbones (ViT-L/14 and EVA02-L-14). Our best submission reached F1 0.92 on the hidden test set. Across all experiments, a consistent 0.026–0.032 AUC validation-to-test gap points to distribution shift from novel generators as the primary challenge.

1. Introduction

The rapid proliferation of generative AI (GANs, diffusion models, and large-scale text-to-image systems) has made synthetic image detection a critical challenge for media forensics. Modern generative models produce images visually indistinguishable from real photographs, yet they carry characteristic statistical fingerprints in spatial and frequency domains that machine learning models can detect [2].

The MediaEval 2026 Synthetic Images Challenge Subtask A frames this as binary classification: given a social media image, classify it as real (label 0) or synthetically generated (label 1). A key challenge, noted in [1], is robustness to real-world transformations (JPEG compression, resizing, and cropping) applied by social media platforms, which can erase the generative artifacts that models rely on.

In this work, we adopt a progressive pipeline that begins with convolutional baselines and culminates in frozen CLIP-based feature extraction. We motivate each architectural transition empirically, using validation AUC as the selection criterion. A recurring theme across all our experiments is the gap between validation and test performance, which we attribute to distribution shift caused by novel generative models present in the test set but absent from any training data. Addressing this shift, rather than maximising validation AUC, is the central challenge of the task.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online

m.junaid.26909@khi.iba.edu.pk (M. Junaid); m.ihsan.27152@khi.iba.edu.pk (M. Ihsan); h.usman.26990@khi.iba.edu.pk (H. Usman); atiftahir@iba.edu.pk (A. Tahir)

CEUR Workshop Proceedings (CEUR-WS.org)

The remainder of this paper is organized as follows. Section 2 describes the dataset and preprocessing setup. Section 3 covers our preliminary CNN experiments, including EfficientNet-B4, ConvNeXt-Tiny, and a frequency-domain dual-stream variant. Section 4 details our CLIP-based approach for both the constrained and open runs. Section 5 presents and analyses test-set results, including the generalization gap across all submissions. Section 6 concludes with a discussion and directions for future work.

2. Dataset and Setup

The ITW-SM validation set contains 9,999 balanced images (5,000 real, 4,999 synthetic) from Instagram, Facebook, LinkedIn, and X. Resolutions range from 512 to 6,000 pixels (median 1,080). We applied a stratified 80/20 train-validation split (seed 42): 7,999 training and 2,000 validation images. For the open run, we added a balanced CIFAKE subset [7] (12,000 images), combining to 21,999 images total. The validation set remained ITW-SM-only for consistent model selection.

For convolutional models we used `LongestMaxSize` (longest side to 224 px) + `PadIfNeeded` (zero-padding to 224×224) + ImageNet normalization, since a direct resize would distort the widely varying aspect ratios. For CLIP models: bicubic resize, center crop, and CLIP normalization statistics, matching the pretraining input distribution.

3. Preliminary CNN Experiments

Before settling on CLIP-based features, we ran several CNN experiments on the ITW-SM split. All used AdamW (weight decay 10^{-4}), CosineAnnealingLR, BCEWithLogitsLoss, mixed-precision training, gradient clipping, and early stopping on validation AUC.

EfficientNet-B4 [6] with strong Albumentations [11] augmentation (JPEG compression simulation, blur/sharpen/motion-blur, GaussNoise, CoarseDropout) reached val AUC 0.9748. Mild augmentation severely overfit (train loss 0.014, val loss 0.86); the JPEG simulation in particular is motivated by social media re-compression that partially erases GAN artifacts.

ConvNeXt-Tiny [5] (28M params, same augmentation, lr=5×10^{-5}) achieved val AUC 0.9906, our best CNN result, with depthwise convolutions and inverted bottleneck blocks producing a more expressive feature hierarchy. We also tested a ConvNeXt + DCT/FFT dual-stream variant motivated by [10], but it underperformed (val AUC 0.9825), likely because ConvNeXt already captures frequency patterns implicitly.

A weighted ensemble of EfficientNet-B4 (30%) and ConvNeXt-Tiny (70%) reached val AUC 0.9984 (F1 0.9864) but only test AUC 0.9675, the largest generalization gap of any submission, confirming that near-perfect validation performance does not transfer.

4. CLIP-Based Approach

To improve generalization, we used frozen CLIP ViT-L/14 [4] (307M params, pretrained on 400M image-text pairs) as a feature extractor. CLIP's diverse pretraining implicitly covers AI-generated imagery from many generative models, providing distribution-invariant representations.

CLIP features are extracted once (7,999×768 train, 2,000×768 val, 10,000×768 test) and cached to Drive (Figure 1). All probe training then operates on cached tensors with batch size 512, reducing epoch time from ∼8 minutes to under 5 seconds and enabling 20–30 epochs within Colab free-tier GPU limits.

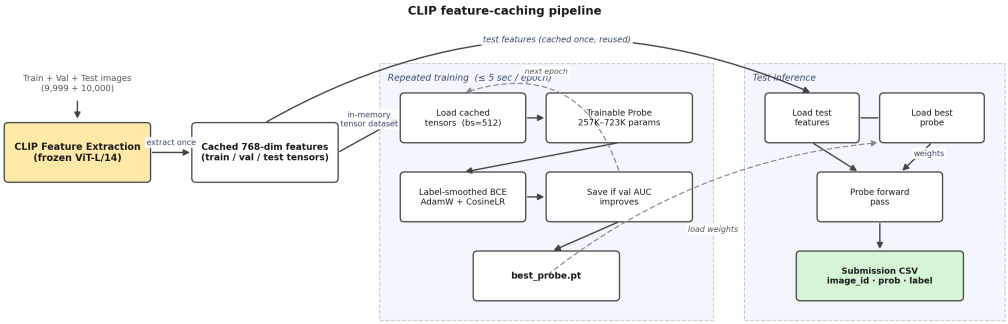

Figure 1: CLIP feature-caching pipeline. Extraction runs once; all probe training and inference use cached tensors.

Constrained Run: Probe Variants. Two probe heads were evaluated on cached ITW-SM features: a linear probe (LayerNorm + 2-layer MLP, 257K params) and a residual probe (with skip connection, 723K params), both trained with AdamW + CosineAnnealingLR and label-smoothed BCE (smoothing=0.05). The linear probe reached val AUC 0.9963 (F1 0.9700); the residual probe reached val AUC 0.9957 (F1 0.9769) and was selected as the constrained-run submission.

Open Run: ViT-L/14 and EVA02-L-14. Both open-run submissions use a deeper 3-layer MLP residual probe with dropout and GELU activations (461K params), trained on the combined ITW-SM + CIFAKE dataset with AdamW + cosine LR, label-smoothed BCE, batch size 512, and early stopping. Color augmentations were deliberately omitted to preserve color-domain discriminative cues; random crop, horizontal flip, and rare grayscale conversion (p=0.05) were applied.

Model 1: CLIP ViT-L/14 uses OpenAI weights via `open_clip` [9], providing a direct comparison against the constrained ViT-L/14 run. *Model 2: EVA02-L-14* [8] is loaded with `merged2b_s4b_b131k` weights and incorporates masked image modeling alongside contrastive pretraining. It is a drop-in replacement requiring no probe changes.

5. Results and Analysis

Test Performance. Table 1 reports test-set performance for both runs. Where the challenge server reported an optimal threshold, we include both default (0.50) and tuned results.

Across all models, false negatives (fake predicted as real) substantially outnumber false positives, consistent with systematic underconfidence in predicting the fake class, which is addressed by lowering the threshold.

Table 1
Test-set results for constrained and open runs. Thr = decision threshold.

Run	Submission	Thr	Acc	P	R	F1	AUC
Const.	CNN Ensemble	0.53	0.885	0.949	0.814	0.876	0.968
Const.	CLIP probe (default)	0.50	0.912	0.949	0.871	0.908	0.969
Const.	CLIP probe (tuned)	0.369	**0.915**	0.927	**0.902**	0.914	**0.969**
Open	ViT-L/14 (default)	0.500	0.912	0.955	0.865	0.908	0.965
Open	ViT-L/14 (tuned)	0.257	—	—	—	0.918	0.965
Open	EVA02-L-14 (default)	0.500	0.911	0.957	0.861	0.907	**0.967**
Open	EVA02-L-14 (tuned)	0.251	—	—	—	**0.921**	**0.967**

The Generalization Gap. Table 2 reveals the most important finding: a remarkably consistent 0.026–0.031 AUC gap between validation and test performance, independent of architecture or training data.

Table 2
Validation vs. test AUC generalization gap.

Model	**Val AUC**	**Test AUC**	**Gap**
CNN Ensemble	0.9984	0.9675	−0.031
CLIP residual probe	0.9957	0.9694	−0.026
ViT-L/14 + CIFAKE	0.9963	0.9645	−0.032
EVA02-L-14 + CIFAKE	0.9946	0.9665	−0.028

The consistency strongly suggests the gap is driven by *distribution shift* (novel generative models in the test set not represented in any training data) rather than any model-specific issue.

Counter-intuitively, the CNN ensemble has the *largest* gap despite the highest validation AUC (0.9984), because CNNs memorize validation-specific generator fingerprints rather than learning distribution-invariant features. Despite lower validation AUC, the CLIP probe achieves higher test AUC (0.9694 vs. 0.9675), consistent with prior work [12] showing CLIP-based features generalize better under distribution shift.

6. Discussion and Outlook

We presented a progressive pipeline for synthetic image detection at MediaEval 2026. Our key findings are: (1) CLIP features generalize better than CNN features despite lower validation AUC; (2) a consistent 0.026–0.032 AUC generalization gap persists across all models, driven by novel generators in the test set; (3) domain-mismatched external data (CIFAKE) hurts generalization rather than helping it; and (4) feature caching makes competitive CLIP training feasible on free-tier GPU hardware. Future directions include fine-tuning the last CLIP transformer blocks and sourcing external data from modern diverse generators that more closely match the test-set distribution.

Acknowledgements

The authors thank the MediaEval 2026 organizers for providing the ITW-SM dataset and evaluation infrastructure.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: Grammar and spelling check, and LaTeX formatting assistance. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] O. Papadopoulou, D. Karageorgiou, C. Koutlis, E. Gavves, H. Mareen, and S. Papadopoulos. Synthetic Images at MediaEval 2026: Advancing Detection of Generative AI in Real-World Online Images. In *Proceedings of MediaEval'26: Multimedia Evaluation Workshop*, Amsterdam, Netherlands and Online, 15–16 June 2026.

[2] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of CVPR*, 2020.

[3] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. On the Detection of Synthetic Images Generated by Diffusion Models. In *Proceedings of ICASSP*, 2023.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of ICML*, 2021.

[5] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *Proceedings of CVPR*, 2022.

[6] M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of ICML*, 2019.

[7] J. J. Bird and A. Lotfi. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *arXiv:2303.14126*, 2023.

[8] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv:2303.15389*, 2023.

[9] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, A. Farhadi, A. Rohrbach, and L. Schmidt. OpenCLIP. *Zenodo*, 2021. https://doi.org/10.5281/zenodo.5143773

[10] T. Dzanic, K. Shah, and F. Witherden. Fourier Spectrum Discrepancies in Deep Network Generated Images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[11] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2), 2020.

[12] Q. Li, A. Ciamarra, R. Caldelli, and S. Berretti. A CLIP-Based Approach for Synthetic Image

Detection under Distribution Shift. In *Proceedings of the MediaEval 2025 Workshop*, Dublin, Ireland and Online, 2025.