Multi-Backbone Frozen Features and OOF-Stacked Ensembling for Synthetic Image Detection

Syed Saad Hasan Emad[1], Abdulrehman Muhammad[1], Hammad Shariq[1], Syed Farhan[1], Syed Muhammad Meesum Abbas[1], Syed Daniyal Hussain[1], Muhammad Atif Tahir[1]

[1] *Institute of Business Administration, Karachi, Pakistan*

Abstract

This paper describes our submission to Subtask A of the MediaEval 2026 Synthetic Image Detection Task. We extract image features using six frozen pre-trained vision backbones and feed the concatenated representation into a stacked ensemble of three classifiers: a wide MLP, a small transformer that processes the features in chunks, and a logistic regression. Ensemble weights and the decision threshold are tuned with Optuna over out-of-fold predictions. On the official hidden test set, our Open Run reaches F1 = 0.9203 (accuracy 92.36%, ROC AUC 0.9695), and our Constrained Run reaches F1 = 0.7111. The 21-point gap between the two runs is the main result of this work: training-data diversity matters more here than the ensemble's architectural complexity.

1 INTRODUCTION

GAN-based models like StyleGAN and ProGAN, and more recently latent diffusion models like Stable Diffusion, can now produce photorealistic images that are difficult to distinguish from photographs. This is a problem for media integrity and for the wider effort against online misinformation. Subtask A of the MediaEval 2026 Synthetic Images Task [5] asks participants to classify images as real or AI-generated under realistic conditions, which include the kind of compression, resizing, and cropping that social media platforms apply to user uploads. Participants submit two runs: a Constrained Run that uses only officially permitted training data, and an Open Run that may use any data the authors document. We submitted both.

Two properties of the task shaped our choices. The test images come from real social media uploads, so our training data and augmentation pipeline had to match that condition rather than the clean datasets typical of academic benchmarks. The set of generative models in use today is also wide and growing, so a detector trained to recognize one family's artefacts is likely to fail on the next. These two points led us away from fine-tuning a single backbone and toward an ensemble of frozen foundation models combined with aggressive domain-aligned augmentation.

2 RELATED WORK

Wang et al. [1] showed that CNN-generated images contain consistent spectral artefacts that a detector can pick up reliably as long as the training and test distributions match. Corvi et al. [2] later documented that this in-distribution success does not carry over once the generator architecture changes, and that the shift from GANs to diffusion has been especially hard on detectors trained on older data.

A separate line of work argues that the response to this distribution shift is to stop fine-tuning the backbone at all. Ojha et al. [3] reported that a frozen CLIP ViT with a linear classifier on top beat end-to-end fine-tuning out of distribution, and Li et al. [6] confirmed this at MediaEval 2025 with a frozen CLIP ViT-L/14 reaching F1 = 0.8870. Several other 2025 submissions independently found that ensembling helps under domain shift. Our submission sits at the intersection of both observations: six frozen backbones with different pretraining paradigms, combined through a stacked ensemble whose weights are tuned with Bayesian optimization.

3 DATA

For the Constrained Run we used only the officially permitted sources. The Wang et al. validation split contributed 4,000 real images and 4,000 ProGAN-generated fakes. The Corvi et al. test split added 15,724 synthetic images

[1]*MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online*
m.a.larson@tudelft.nl (M. Larson); gareth.jones@computing.dcu.ie (G. Jones); bionescu@imag.pub.ro (B. Ionescu)

covering StyleGAN, BigGAN, diffusion, and transformer-based generators. We sampled 15,724 real images at random from COCO 2017 to balance the real side. The final training set had 19,724 real and 19,724 fake images.

The Open Run reused the Constrained set and added two external sources. We took 9,000 synthetic images from Synthbuster, which gave us coverage of modern diffusion outputs that the Constrained pool was thin on. We added 8,111 high-resolution real photographs from RAISE, converted from RAW to JPEG. These bring sensor-level characteristics of genuine camera capture into the training set, which we expected to help the model separate photographs from synthetic images at the pixel-statistics level.

Preprocessing was identical for both runs. Images were resized so the shorter side was 256 pixels and saved as JPEG. At training time the augmentation pipeline simulated the kinds of transformations a social media upload tends to go through: random resized crops to 224, horizontal and vertical flips, JPEG compression at quality 30 to 95, color jitter, random grayscale, AutoContrast, Equalize, Gaussian blur, rotation up to ±20°, and Random Erasing. At evaluation time we resized to 256, took a deterministic center crop of 224, and normalized with ImageNet statistics.

4 APPROACH

4.1 Feature extraction

Fine-tuning the backbones was not realistic on our hardware, and we were also concerned that fine-tuning on the available training data would teach the model to recognize artefacts specific to that data rather than generalize. We therefore kept all six backbones frozen and used them only as feature extractors. The six models, taken from the `timm` library, cover three different pretraining paradigms: DINOv3 self-supervised pretraining (ConvNeXt Base, ConvNeXt Large, ViT-B/16, ViT-L/16), Masked Autoencoder pretraining [8] (ViT-L/16), and CLIP contrastive pretraining [7] (ViT-L/14). Their output dimensions range from 768 to 1,536. The full set is listed in Table 1.

Table 1: Frozen feature extractors used in our pipeline

Model	Architecture	Pretraining	Output
convnext_large.dinov3_lvd1689m	ConvNeXt Large	DINOv3 Self-Supervised	1536
convnext_base.dinov3_lvd1689m	ConvNeXt Base	DINOv3 Self-Supervised	1024
vit_large_patch16_224.mae	ViT-L/16	Masked Autoencoder (MAE)	1024
vit_large_patch16_dinov3.lvd1689m	ViT-L/16	DINOv3 Self-Supervised	1024
vit_base_patch16_dinov3.lvd1689m	ViT-B/16	DINOv3 Self-Supervised	768
vit_large_patch14_clip_224.laion2b_ft_in12k	ViT-L/14	CLIP Contrastive	1024

We concatenated the embeddings from all six models into a single vector of roughly 6,400 dimensions per image. The choice of paradigms was intentional. DINOv3 and MAE pretraining objectives both push the backbone to model local pixel statistics and structural consistency, so the features they produce are well-suited to picking up low-level synthesis artefacts. CLIP's contrastive image-text objective produces features that encode scene-level semantic content, which lets the detector pick up the high-level semantic anomalies and cross-modal alignment cues that distinguish a generated image from a photograph of a real scene. We expected the two kinds of signal to be complementary. Extraction ran in parallel across two NVIDIA T4 GPUs on Kaggle using Python's ThreadPoolExecutor, and we saved the resulting embeddings as NumPy arrays so they could be reused across runs without recomputation.

4.2 Stacked ensemble

We trained three base classifiers on the concatenated features. The first is a wide MLP with two hidden layers of 1,024 and 512 units, with BatchNorm, ReLU activations, and Dropout of 0.2, trained with AdamW and binary cross-entropy. The

second is a small transformer encoder we refer to as ChunkTransformerLite. It splits the 6,400-dimensional feature vector into chunks of 100 elements, projects each chunk to 64 dimensions, runs two layers of 4-head self-attention, mean-pools the result, and applies a linear classification head. The motivation for this classifier was to let the model discover useful interactions between features that originated from different backbones. The third classifier is an L2-regularized logistic regression on StandardScaler-normalized features, which gives us a stable, well-calibrated linear baseline.

To avoid leakage when learning the ensemble weights, we used 10-fold stratified cross-validation and collected out-of-fold predictions from all three base classifiers into an N × 3 matrix. Before any weighting step, we converted each model's predictions to percentile ranks using `scipy.stats.rankdata`. This normalizes the output scales of the three classifiers, which would otherwise sit on quite different calibration curves.

4.3 Weight and threshold optimization

We tuned the ensemble weights and the decision threshold jointly rather than averaging the three classifiers uniformly. The tuning ran in three stages. A coarse grid search over the weight pair (w_wide, w_trans), combined with a threshold sweep over 0.1 to 0.9 in 181 steps, gave us an initial picture of the search space. The top 20 (weights, threshold) combinations by OOF F1 were retained as warm-start hints. We then ran an Optuna TPE search with 1,500 trials and 50 startup trials, seeded by those 20 grid points as warm starts. Finally, a ±0.05 local search around the Optuna best point, paired with a 101-step threshold sweep, confirmed and slightly refined the result. The objective at every stage was out-of-fold F1.

The final Open Run weights came out at 0.834 for the WideMLP, 0.166 for the logistic regression, and zero for ChunkTransformerLite, with a decision threshold of 0.324. We return to the zero weight on the transformer in the analysis.

5 RESULTS

The **Open Run** reached **F1 = 0.9203** with accuracy 0.9236, precision 0.9620, recall 0.8820, ROC AUC 0.9695, and average precision 0.9748. Of the 5,000 real and 5,000 fake images in the hidden test set, this corresponds to 4,826 real images correctly classified (174 false positives) and 4,410 fake images correctly classified (590 false negatives).

The **Constrained Run** reached **F1 = 0.7111** with accuracy 0.6648, precision 0.6248, recall 0.8250, ROC AUC 0.7344, and average precision 0.7161. The confusion matrix here is much less clean: 2,523 real images correctly classified and 2,477 false positives, alongside 4,125 fake images correctly classified and 875 false negatives.

Table 2: Open run Confusion Matrix

	Pred Real	Pred Fake	Total
Real	4826	174	5000
Fake	590	4410	5000

Table 3: Constrained run Confusion Matrix

	Pred Real	Pred Fake	Total
Real	2523	2477	5000
Fake	875	4125	5000

6 ANALYSIS

The Constrained Run's precision-recall split is the most striking thing in our results. Precision was 0.625 against a recall of 0.825, meaning 2,477 out of 5,000 real images were flagged as synthetic. The Constrained training pool is heavy on GAN-era fakes and is narrow in terms of real-image categories, so our reading is that the model learned to associate certain spectral signatures with the synthetic class, and that those same signatures also appear in some real photographs that have unusual textures or high-frequency content. Adding Synthbuster diffusion samples and high-quality RAISE photographs to the Open Run dropped false positives from 2,477 to 174. This is fairly direct evidence that under data restriction, the bottleneck is real-image diversity at least as much as it is fake-image coverage.

A second result that surprised us is that ChunkTransformerLite ended up with a weight of exactly zero in the final ensemble. The optimizer was free to assign it any nonzero weight and chose not to. Our interpretation is that the frozen DINOv3 and CLIP embeddings are already organized into directions that a dense MLP can separate, so adding chunk-level self-attention does not give the ensemble any new information to work with. This lines up

with what Li et al. [6] reported: under distribution shift, a simple head on a strong frozen feature set tends to do better than more elaborate architectures.

The Open Run still misses 590 fakes. We did not have time to do a per-generator breakdown of these failures, but our hypothesis is that they concentrate on modern photorealistic diffusion outputs that are close to indistinguishable from RAISE-quality photographs at the level of frozen-backbone features. A failure analysis split by generator would test this directly, and is the most useful next step.

A separate factor worth singling out is the augmentation pipeline. The training images, both before and after we added the open-run sources, are largely clean; the hidden test images are not. Our augmentation choices were designed to close that gap: heavy random JPEG compression in the quality 30 to 95 range, Gaussian blur, random crops and rotations, color jitter, AutoContrast, Equalize, and Random Erasing, all chosen to approximate the kinds of transformations a social media platform applies to user uploads. We did not have time to ablate individual augmentations, but the size of the distributional gap between training and test conditions, combined with the strong Open Run F1, suggests this part of the pipeline carried real weight.

The 21-point F1 gap between the Constrained (0.7111) and Open (0.9203) runs is the cleanest summary of what we learned from this task. For synthetic image detection under in-the-wild conditions, training-data diversity and the augmentation pipeline are the two factors that mattered most in our experiments, ahead of how deep or how attention-heavy the ensemble is.

7 CONCLUSION & FUTURE WORKS

Our submission combined six frozen pre-trained backbones, an OOF-stacked ensemble of three classifiers, and Optuna-tuned ensemble weights. It reached F1 = 0.9203 on the official MediaEval 2026 hidden test set. The comparison between the Constrained and Open Runs makes a fairly clean argument that training-data diversity is the single most important factor for cross-domain generalization on this task. Several directions look worth pursuing next: a systematic ablation of the augmentation pipeline to see which transformations actually matter, a per-generator analysis of the Open Run's false negatives, the addition of frequency-domain features such as DCT or FFT coefficients as auxiliary channels, and a semi-supervised approach that takes advantage of the unlabeled test images.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) for drafting content, paraphrasing and rewording. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

REFERENCES

[1] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A. CNN-generated images are surprisingly easy to spot... for now. CVPR, 2020.

[2] Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L. On the detection of synthetic images generated by diffusion models. ICASSP, 2023.

[3] Ojha, U., Li, Y., Lee, Y.J. Towards universal fake image detectors that generalize across generative models. CVPR, 2023.

[4] Papadopoulou, O., Schinas, M., Corvi, R., Karageorgiou, D., Koutlis, C., Guillaro, F., Gavves, E., Mareen, H., Verdoliva, L., Papadopoulos, S. Synthetic Images at MediaEval 2025: Advancing Detection of Generative AI in Real-World Online Images. In Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[5] Papadopoulou, O., Karageorgiou, D., Koutlis, C., Gavves, E., Mareen, H., Papadopoulos, S. 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.

[6] Li, Q., Ciamarra, A., Caldelli, R., Berretti, S. A CLIP-based Approach for Synthetic Image Detection under Distribution Shift. MediaEval 2025 Workshop.

[7] Radford, A., et al. Learning Transferable Visual Models from Natural Language Supervision. ICML, 2021.

[8] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. Masked Autoencoders Are Scalable Vision Learners. CVPR, 2022.