Detecting Synthetic Images with Frozen DINOv3 Layer Features: A Layer-Depth Study

Vedant Gupta[1,*]

[1]*Indian Institute of Technology Hyderabad, India*

Abstract

ResolveNow submits to Task A of the MediaEval 2026 Synthetic Images challenge, binary classification of real against synthetic images drawn from in-the-wild sources. The detector reads frozen patch tokens from DINOv3 ViT-L and trains a three-layer Transformer head over the concatenation of two backbone layers, with no pooling at any stage. A sweep across single layers and pairs separates DINOv3 depth into three regimes. Layers 1 through 6 give below-chance AUC (0.339 to 0.429), layers 8 through 12 sit near random (0.505 to 0.541), and layers 16 through 24 carry the discriminative signal (0.686 to 0.917), with the sharpest transition between layer 12 and layer 16. Pairing any layer with layer 24 reaches AUC 0.919 to 0.937. We select the layer 12 and layer 24 pair for both runs. It scores the highest F1 of any configuration on the sweep subset (0.793), and F1 is the task metric. On the 10k test set, the constrained model trained on GAN and latent-diffusion data reaches F1 0.718 at recall 0.593, missing 40.7 percent of synthetic images. Continuing training on the provided validation images for the open run raises recall to 0.862 and F1 to 0.902. The gain is almost entirely recovered recall. Both runs submit at threshold 0.5, while the F1-optimal thresholds fall at 0.001 and 0.097, which places most predicted probabilities near zero.

1. Introduction

We participate in Task A of the MediaEval 2026 Synthetic Images challenge. The task asks for binary classification of images as real or synthetic and ranks systems by F1 [1]. It requires one constrained run trained only on the provided data and one open run that may draw on any source. Test images arrive from social media after compression, resizing, and cropping. These transformations erase the low-level forensic traces that frequency-domain and noise-residual detectors read. Detectors trained on laboratory data fail under these conditions.

The backbone choice decides what the detector reads, and our first choice failed in a way that shaped the design. A classifier on Qwen2.5-VL patch features reached validation AUC 1.0 by reading the model's variable patch count rather than image content. DINOv3 returns 196 patch tokens for any image at fixed resolution, which removes that artifact without pooling [2].

Two questions follow from this design. A frozen backbone exposes every block, so the first is which DINOv3 layers carry the signal. The second is whether a detector trained on laboratory data transfers to in-the-wild images. Discriminative signal concentrates in the late layers. Concatenating a middle layer with the final layer recovers most of the available performance. A constrained detector trained only on GAN and latent-diffusion data under-detects synthetic images outside its training distribution, and continued training on the provided validation set recovers the missing recall.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
ce23btech11059@iith.ac.in (V. Gupta)

2. Related Work

Frozen foundation-model features are the dominant paradigm for synthetic image detection. Ojha et al. [3] show that a linear probe on CLIP features generalizes across generators that a trained CNN detector misses. RINE extends this by reading intermediate CLIP blocks rather than the final representation, on the premise that different blocks encode different generative traces [4]. TextureCrop addresses the resolution problem by cropping high-texture regions before classification [5]. Prior MediaEval submissions established CLIP and SigLIP probes as baselines for this task [6]. We build on this line of work with a frozen DINOv3 backbone, a small attention head in place of the linear probe, and a sweep over which backbone layers feed it.

3. Approach

Our first detector read patch features from Qwen2.5-VL [7], a vision-language model whose dynamic tokenizer emits a different patch count per image. Every synthetic image in our data resolved to 400 patches while real images produced between 520 and 2116. A classifier over these variable-length sequences reached validation AUC 1.0, but it read sequence length rather than image content. We confirmed the shortcut by passing solid-color and resized real images through the trained model so that the same content produced a different patch count, and the prediction tracked the count rather than the content. Forcing a fixed length through pooling would remove the artifact, but pooling discards the per-patch structure a forensic detector depends on. This failure sets the central constraint, no pooling at any stage, and points to a backbone that returns a fixed token count on its own.

The architecture follows from that constraint and one open question, which layers to read. The first stage extracts patch tokens from a frozen DINOv3 ViT-L encoder (dinov3-vitl16-pretrain-lvd1689m). The DINOv3 processor resizes each image to 224 by 224 and produces 201 tokens, from which we drop the class token and four register tokens to keep 196 patch tokens of dimension 1024.

The second stage sweeps across layers. Each layer supplies a tensor of shape 196 by 1024, and concatenating two of them along the feature axis produces a tensor of shape 196 by 2048. Concatenation along features rather than along the token axis preserves spatial correspondence between the two layers, so each patch keeps a single position with a 2048-dimensional descriptor.

A three-layer Transformer head consumes the resulting tensor, whether it comes from a single layer or a concatenated pair. A LayerNorm over the input feeds a linear projection to 256 dimensions. A learnable class token prepends to the patch sequence, and a pre-norm encoder with four heads and feed-forward width 1024 processes the sequence under a padding mask. The class-token output passes through a LayerNorm, dropout at 0.1, and a linear layer to a single logit. The head holds 2.9M parameters against the frozen backbone. Training uses binary cross-entropy with a positive-class weight set to the real-to-fake ratio, AdamW at weight decay 0.01, gradient clipping at norm 1.0, and a cosine schedule. The constrained run trains for 30 epochs at learning rate 3e-4. The open run restores the constrained checkpoint and continues for 10 epochs at 1e-4.

The constrained run trains only on the official data, the GAN benchmark of Wang et al. [8] (StyleGAN2, BigGAN, ProGAN) and the latent-diffusion and COCO/SUN collection of Corvi et al. [9]. The open run adds the validation set images. The validation set, 10k images split evenly between real and synthetic, supports development and the layer sweep. The organizers score the two submitted runs on a held-out 10k test set, which supplies the run metrics reported here.

Table 1
Layer sweep on the 1k validation subset, 10 epochs at seed 42.

Config	AUC	F1	Acc
L1	0.339	0.115	0.460
L4	0.429	0.090	0.496
L6	0.390	0.027	0.495
L8	0.505	0.004	0.498
L12	0.541	0.043	0.507
L16	0.740	0.207	0.548
L18	0.686	0.153	0.533
L20	0.766	0.222	0.551
L24	0.917	0.763	0.803
L1 + L24	0.926	0.585	0.704
L4 + L24	0.926	0.690	0.760
L6 + L24	0.934	0.705	0.769
L8 + L20	0.726	0.047	0.511
L12 + L24	0.930	**0.793**	0.824
L16 + L20	0.767	0.163	0.537
L16 + L24	**0.937**	0.749	0.796
L18 + L24	0.919	0.642	0.732

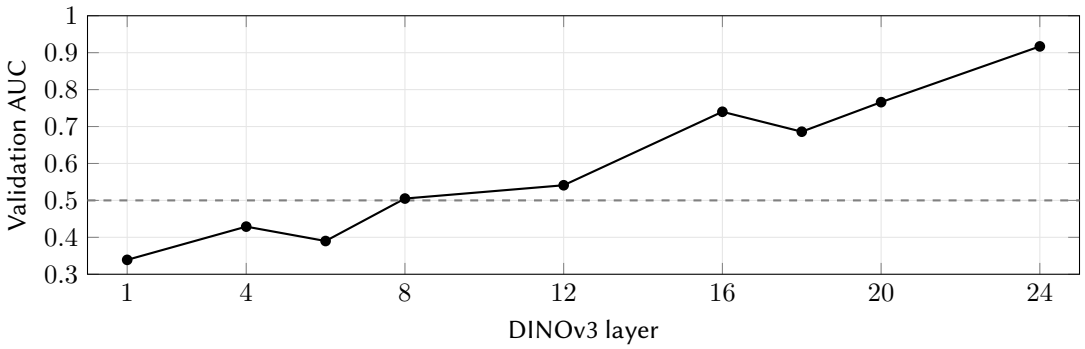

Figure 1: Single-layer validation AUC by DINOv3 depth on the 1k subset. The dashed line marks chance. Layers 1 to 6 fall below it, layers 8 to 12 sit near it, and layers 16 to 24 rise toward 0.917, with the sharpest step between layer 12 and layer 16.

4. Results and Analysis

Layer depth tracks detection quality, and the sweep splits DINOv3 into three regimes on a single 1k subset at one seed (Table 1, Figure 1). We report these as observations from one run per configuration, a suggestive ordering rather than a measured effect with error bars. Early layers fall below chance, from AUC 0.339 at L1 to 0.429 at L4, and the middle layers sit near random. Discriminative signal appears between layer 12 and layer 16, a 20-point AUC step, and the strongest content sits in the deepest third of the network.

A single late layer carries most of that content. A second layer sharpens it. Every pair that includes layer 24 lands between 0.919 and 0.937, while two weak layers paired together stay low. L16+L24 takes the top AUC and L12+L24 the top F1. We submit L12+L24 because the task scores F1.

The selected detector, trained on laboratory data, under-detects on the test set (Table 2). The

Table 2
The two submitted runs on the 10k test set at threshold 0.5, 5000 real and 5000 synthetic. TN and FP are real images, FN and TP synthetic.

	Counts				Metrics					
Run	TN	FP	FN	TP	Acc	Prec	Rec	F1	AUC	AP
Constrained	4710	290	2037	2963	0.767	0.911	0.593	0.718	0.799	0.792
Open	4749	251	690	4310	0.906	0.945	0.862	0.902	0.964	0.969

error is one-sided. The constrained model rarely misfires on real photographs but leaves a large share of synthetic images undetected. It recognizes the GAN and latent-diffusion traces it trained on and abstains on the generators it never saw.

Continued training on the provided validation images recovers the missing detections. Undetected synthetic images drop sharply while false alarms on real images hold near their prior level, so recall rises and precision rises with it. The gain is recovered recall, not a precision-recall trade. The open run trains on the validation images and is scored on the held-out test set, so these are held-out numbers.

The submitted threshold of 0.5 sits above the F1 optimum for both runs. The F1-maximizing threshold is 0.001 for the constrained run, lifting its F1 to 0.734, and 0.097 for the open run, lifting it to 0.908. The detector rarely emits a high synthetic probability even when correct, so 0.5 costs recall. We submitted 0.5 because the test set carried no labels for threshold fitting.

5. Discussion and Outlook

The recall gap between the two runs locates the failure of laboratory-only training. The constrained model does not produce false alarms on real photographs, since its precision stays above 0.9 and rises after adaptation. It fails by under-detection, returning low synthetic probabilities for generators outside its training distribution. Adaptation on the provided validation set raises the probability the head assigns to those generators without trading away real-image precision.

Early-layer features carry a related risk in a sharper form. A head on a single early layer scores below chance on the validation subset, where L6 reaches AUC 0.390 against L24 at 0.917, so the layer is unreliable on its own. The selected pair leans on the final layer for this reason.

The MediaEval question on training-data influence sets the next steps. Inspecting the generator composition of the 2037 constrained false negatives would identify which families a GAN-plus-latent-diffusion set leaves uncovered. Threshold calibration on labeled in-distribution data would recover the F1 that the default 0.5 leaves unused.

Acknowledgments

We built this system with DINOv3 [2]. We thank the MediaEval Synthetic Images task organizers for the data and the evaluation.

Declaration on Generative AI

During the preparation of this work, the author used a large language model (Anthropic Claude) to refine prose, check grammar and style, and assist with LaTeX formatting, and to help organize and cross-check the reported results against the source code and experiment logs. No figures or

data were generated by a generative AI system. All results come from the experiments described in the paper. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] O. Papadopoulou, D. Karageorgiou, C. Koutlis, E. Gavves, H. Mareen, 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, 2026.

[2] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, P. Bojanowski, DINOv3, arXiv preprint arXiv:2508.10104 (2025).

[3] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[4] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, in: Proceedings of the European Conference on Computer Vision (ECCV), Springer, 2024, pp. 394–411.

[5] D. Konstantinidou, C. Koutlis, S. Papadopoulos, Texturecrop: Enhancing synthetic image detection through texture-based cropping, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 2025, pp. 1459–1468.

[6] O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, S. Papadopoulos, 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, pp. 25–26.

[7] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-VL technical report, arXiv preprint arXiv:2502.13923 (2025).

[8] 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 (CVPR), 2020.

[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 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.