Foundation Models for Face Presentation Attack Detection: A Unified Linear-Probing Benchmark

Peter Lorenz, Anjith George, Sebastien Marcel
Idiap Research Institute
Rue Marconi 19, Martigny, Switzerland.
{peter.lorenz, anjith.george, sebastien.marcel}@idiap.ch

Abstract

Face presentation attack detection (PAD) remains challenging under cross-dataset evaluation, where domain shift degrades models trained on a single dataset. The scarcity of large-scale labeled data motivates adapting pretrained vision models rather than training task-specific architectures from scratch, raising a fundamental question: do general-purpose vision foundation models encode PAD-relevant information accessible with minimal task-specific training? To investigate, we systematically evaluate 24 frozen encoders, including self-supervised Vision Transformers, vision–language encoders, and supervised CNNs, using a unified linear-probing protocol on the MCIO benchmark (MSU-MFSD, CASIA-FASD, Replay-Attack, OULU-NPU). The backbone remains fixed, and only a lightweight linear head is trained to isolate the PAD information already present in the pretrained representation. We report intra- and cross-dataset performance, along with accuracy–compute trade-offs, relative to two specialist PAD baselines. Results show that frozen foundation-model representations can support strong intra-dataset PAD performance with only a linear classifier, but this performance does not reliably transfer across datasets. Model scale is beneficial within several families, although the effect is not monotonic and is strongly mediated by architecture and pretraining. InternViT-6B achieves the lowest mean intra-dataset error, whereas CLIP ViT-B/32 offers the most favorable cross-dataset transfer–compute trade-off among the evaluated probes. These findings suggest that while pretrained representations contain PAD-relevant information, explicit adaptation remains necessary to address domain shift. Code and evaluation protocols will be publicly released. gitlab.idiap.ch/paper/foundationpad.

1. Introduction

Face recognition (FR) systems have improved substantially with deep learning [17, 7], yet remain vulnerable to presentation attacks (PAs) [39], such as printed photographs, replayed videos, and related spoofs presented to impersonate an enrolled identity and gain unauthorised access [66, 20, 41] or avoid recognition (obfuscation). Presentation attack detection (PAD) distinguishes bonafide presentations from attack samples before verification [37, 61, 60, 19]. Specialist deep neural network (DNN) approaches to PAD can achieve strong performance when trained and tested on the same dataset, but degrade under cross-dataset shifts in sensors, illumination, and attack types [44, 45, 30, 52, 57, 18]. The four MCIO PAD datasets (MSU-MFSD, CASIA-FASD, Replay-Attack, OULU-NPU) provide far fewer labelled samples than modern vision pretraining datasets such as ImageNet [16], limiting what end-to-end PAD models can learn from scratch [21, 18].

One approach to addressing data scarcity and improving PAD detection is to leverage large-scale vision encoders and multimodal information. Human learning is naturally multimodal, drawing on multiple sensory modalities to better understand and process new information. Vision foundation models (VFMs) and multimodal large language models (MLLMs) pretrained on large-scale image or image–text data offer an alternative: generic representations that may already encode cues useful for spoof detection [43, 31, 38, 12].

Recent works explore the use of foundation models (FMs) for PAD through multimodal architectures (e.g., FLIP [47],

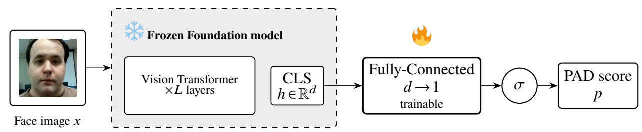


Figure 1. Overview of the linear probing for binary detection: A frozen pre-trained ViT vision encoder maps the face image to a CLS embedding h via L Vision Transformer (ViT) layers. In probing, only the fully-connected layer (FC) is trainable cf. Table 3), and has the input shape of the CLS embedding d mapping to 1. The probability of a sample being bonafide is $p = \sigma(z)$, where z is the linear output (logit) of the FC, and the sigmoid function $\sigma(\cdot)$ converts the logit to a probability for binary classification. The target label is 1 for bonafide samples and 0 for attacked samples.

I-FAS [64], InstructFLIP [34]), fine-tuning (e.g., with LoRA [27] on CLIP [43] in FoundPAD [40], and on DINOv2 with Registers [22]), and self-supervised pre-training without labels, i.e., FSFM [51]. By modifying the pretrained weights of the VFM, these methods primarily assess adaptation performance rather than the general-purpose generalization properties of foundation models (FMs).

Recently, Gonzalez-Soler et al. investigated VFMs under zero-shot and linear-probing settings for PAD [24]. In the linear-probing setting, they add only a trainable classification layer without altering the pretrained backbone. However, their experiments are limited to CLIP and DINO [38]. Building on this idea, we systematically assess whether general-purpose foundation models (FMs) or other large-scale vision encoders encode sufficient PAD-relevant representations.

The research question is: under a fixed linear-probing protocol on the PAD dataset, which general-purpose foundation model (FM) already separates bonafide from attacks intra-dataset, which of that performance survives cross-dataset transfer, and at what compute cost?

Keeping every backbone frozen isolates pretraining strategy and model scale from task-specific adaptation recipes such as LoRA or full fine-tuning. We benchmark 24 pretrained vision encoders, including vision-only models and vision encoders from VLMs and MLLMs. Two PAD specialist models (DeepPixBiS [23] and FSFM-FAS [51]) serve as specialist baselines spanning classical convolutional neural network (CNN) PAD and recent ViT pretraining.
Our main contributions are

- a unified frozen-backbone linear-probing benchmark across a broad set of vision encoders (including vision encoders from VLMs and MLLMs) on standard PAD datasets.
- Besides intra-dataset evaluation, we also conduct cross-dataset evaluation and rank the models according to their performance in both settings.
- We also compare against representative specialist PAD baselines to contextualize when generic FMs justify their compute footprint.

2. Related Work

Cross-dataset PAD and unseen-attack generalization. PAD methods exploit texture, colour, and frequency cues that distinguish live skin from printed or replayed media [37, 32, 19]. Deep Neural Network (DNN) based models excel intra-dataset but degrade under cross-dataset shift [52, 57, 18]. Cross-dataset evaluation trains on one corpus and tests on another, exposing sensor and illumination shift [36, 18]. Prior work addresses this through domain adaptation (DA), which aligns source and target feature spaces but requires target-domain data at training time [32, 50], or domain generalization (DG), which trains on multiple labelled sources using adversarial or meta-learning objectives [44, 45, 30, 13].

Unseen-attack protocols hold out against presentation attacks (PAs) at test time. Leave-one-out (LOO; train on three MCIO corpora, test on the fourth) MICO [44] and limited source domain (LSD) protocols [22] (train on 2 and evaluate on the remainder) measure generalisation when multiple labelled sources are available at training time. Auxiliary datasets, such as CelebA-Spoof [66] and SynthA-Spoof [21], are injected during training to counterfeit data scarcity and to improve generalization capabilities [24].

DNN PAD methods. Pixel-wise supervision [37] and depth-aware architectures [19] train compact CNNs end-to-end on MCIO-scale data and can reach strong intra-dataset ACER. Instead of predicting a single live/spoof label or a dense depth map, the method predicts a pixel-wise binary map: Bona fide image: every location in the supervision map is labeled as 1. Attack image: every location is labeled as 0. The network therefore learns local spoof artifacts through dense supervision while avoiding the complexity of depth estimation. Frequency-domain and multi-scale designs further target print-and-replay artefacts [20]. Patch-based formulations such as PatchNet [49] treat spoof detection as fine-grained local discrimination. These networks remain useful compact deployment baselines because they are lightweight and dataset-tuned, but they often overfit acquisition conditions and degrade under cross-domain shift [57, 18].

Foundation Models for PAD. Vision foundation models (VFMs) pretrained on web-scale image or image–text data offer an alternative to training PAD networks from scratch on small MCIO datasets [43, 38]. FoundPAD [40] adapts CLIP to PAD with Low-Rank Adaptation (LoRA) [27], updating a small subset of backbone weights while preserving most pretrained knowledge. FSFM [51] pretrains a vision transformer CLIP-B/16 backbone with face-spoofing-specific self-supervised learning (SSL) before downstream face anti-spoofing (FAS) fine-tuning. Gonzalez-Soler *et al.* [24] study zero-shot PAD with pretrained ViTs on the SiW-Mv2 dataset [25], removing PAD-specific training entirely.

Later work extends this to ViTs through patch-level augmentation and patch-wise losses [53], a line that recent VFM studies [22] combine with large-scale self-supervised learning (SSL) backbones. SSL is a technique in which a model automatically generates its own training labels from the raw data, rather than relying on manual human annotations.

Feng *et al.* [22] compare frozen backbones and find that self-supervised ViTs outperform supervised CNNs and ImageNet ViTs, with DINOv2 augmented with register tokens [15] yielding the most consistent cross-domain accuracy. They further show that PAD-specific adaptation, including Face Anti-Spoofing Data Augmentation (FAS-Aug) [8], patch-wise data augmentation (PDA) [53], and attention-weighted patch loss (APL) [53], on a fine-tuned DINOv2 backbone closes much of the gap to large vision–language FAS systems at a fraction of the parameter count. Other recent cross-dataset methods explore prompt learning [35], frequency-domain modelling [9], domain–modality alignment [59], and optimal-transport adaptation [33]. Together, these approaches map a spectrum from zero-shot inference [24], through lightweight probing, task-adaptive LoRA [40], to heavily engineered cross-domain training [35, 9, 33, 51, 59, 22].

3. Approach

3.1. Presentation Attack Detection Baseline

To contextualize the frozen-probing results, we compare them with two end-to-end specialized PAD models trained or fine-tuned directly for the task. These specialist baselines span classical CNN-based supervision and SSL-pretrained ViT fine-tuning. DeepPixBiS [23] is a strong DNN approach trained with pixel-wise supervision.

FSFM-FAS [51] is a ViT-B/16 fine-tuned after SSL pre-training on the VGGFace2 dataset [10]. Wang et al. report that FSFM-FAS outperforms the few-shot FAS approach of Huang et al. [28] on their benchmarks [51].

3.2. Frozen FM Backbone with Linear Probing

We evaluate frozen vision encoders with linear probing (e.g., for ViT see Fig. 1). Each cropped face image x is passed through a pretrained vision backbone whose weights remain fixed throughout training and testing. The frozen backbone outputs a global image embedding $h \in \mathbb{R}^d$, whose extraction rule depends on the encoder family (CLS token, post-layer-norm CLS, global average pooling, projected CLIP image embedding, or mean-pooled patch tokens); see the Feature column in table 3. Only a single fully connected layer maps h to a scalar logit $z = w^\top h + b$.Training minimizes binary cross-entropy on z with bonafide labels y=1 and attack labels y=0. At inference, the bonafide probability is computed as: $p = \sigma(z)$.

3.3. Vision Encoders

Table 1 lists the 24 frozen backbones grouped by family and scale. ViT-style encoders (CLIP, DINOv2/v3, BEiT, SigLIP, ViT-MAE/MSN, InternViT) use the final CLS token; Qwen2 uses mean pooling over patch tokens; ConvNeXt-V2 uses global average pooling.

Multimodal LLM Vision Encoders. Multimodal large language models (MLLMs) integrate vision and language within a unified architecture, formally $\mathcal{M}_{\text{MLLM}} = \mathcal{V} \oplus \mathcal{L} \oplus \mathcal{O}$, where $\mathcal{V}$ denotes the vision encoder, $\mathcal{L}$ the language model, and $\mathcal{O}$ optional other modalities. InternVL [12] follows a ViT-MLP-LLM paradigm with InternViT (300M, 6B parameters) as the vision tower and Qwen3 [58] as the language model; a multi-layer projector bridges visual representations to the language embedding space. The full pipeline is optimized end-to-end on large image-text datasets through contrastive alignment (matching image-caption pairs) and language-model loss (generating coherent text from projected visual tokens). QwenVL's vision encoder [4] spans 0.5B–72B parameters and employs a mixture-of-experts design that activates a subset of experts per token for efficient computation.

Table 1. Frozen model backbone zoo by family for linear probing. Within the family, the models only differ in parameter size (in Billions).

Family	Variants	Params (B)	Supervision	Pretraining objective
ConvNeXt-V2	base, huge	0.09, 0.66	Supervised	ImageNet classification
DINOv2	base, giant, w/reg.	0.09–1.14	SSL	Self-distillation
DINOv3	base, giant	0.09, 0.84	SSL	Self-distillation
BEiT	base, large	0.09, 0.30	SSL	Masked image modeling
ViT-MAE	base, huge	0.09, 0.63	SSL	Masked autoencoding
ViT-MSN	base, large	0.09, 0.30	SSL	Masked neighbours
CLIP	B/32, L/14, H/14	0.09–0.63	Vision–lang.	Img–txt contrastive
SigLIP	base, large	0.09, 0.32	Vision–lang.	Sigmoid contrastive
EVA-CLIP	8B	8.22	Vision–lang.	Img–txt contrastive
Qwen2.5-VL	3B, 7B	0.67, 0.68	Multimodal	Img–txt contrastive
InternViT	300M, 6B	0.30, 5.54	Multimodal	VLM vision tower

SSL: self-supervised (no manual class labels). Supervised: ImageNet-1K classification (ConvNeXt-V2 CNN). Vision–lang.: image–text pair pretraining. Qwen2.5-VL variants extract the frozen ViT from 3B/7B VLM checkpoints; both towers are ~0.67 B. Native input 224 px except SigLIP-large (256 px) and InternViT (448 px).

Vision-language Contrastive Models. CLIP [43] (B/32, L/14, H/14) learns aligned image-text representations through contrastive learning on 400 million web-crawled image-caption pairs, enabling zero-shot transfer to tasks for which it was not explicitly trained. The model variants differ in scale (Base, Large, Huge) and patch size (32 vs. 14 pixels), where larger patches reduce computational cost while potentially discarding fine-grained details. SigLIP [63] (Base, Large) replaces the softmax in CLIP with a sigmoid loss, allowing independent processing of each image-text pair and more efficient batch-size scaling, which is particularly beneficial for large-scale training. EVA-CLIP-8B [48] pushes CLIP to larger scale by initializing the vision encoder with masked-image pretraining (similar to BERT for text) before contrastive fine-tuning on web-scale data.

Self-supervised Vision Encoders. BEiT [5] (Base, Large) performs masked image modeling by predicting discrete visual tokens obtained from a pre-trained discrete VAE for masked patches, excelling at semantic understanding tasks. DINOv2 [38] combines self-distillation (learning without labels) with iBOT [67] (masked modeling in latent space) and is trained on a curated 142M-image dataset (LVD-142M), built by augmenting ImageNet-22k [16] and Google Landmarks [55] with visually similar web images. DINOv3 [46] addresses scaling limitations of its predecessor with a modified training recipe (constant hyperparameters replacing cosine schedules) and Axial RoPE for positional embeddings, resulting in improved spatial precision and robustness; it is trained on the larger, more diverse LVD-1689M dataset. MSN [3] learns by matching heavily masked views to prototypes (learned visual centers) derived from unmasked views, though it performs near chance in our setting, suggesting that not every self-supervised objective yields PAD-relevant features.

Generative masked image modeling. ViT-MAE [26] (Base, Huge) directly reconstructs RGB pixel values of masked patches without relying on a discrete tokenizer, learning rich representations through a simple and scalable reconstruction objective.

Modern CNNs. ConvNeXt-V2 [56] (Base, Huge) modernizes the classic ResNet architecture with transformer-inspired design elements such as large-kernel depthwise convolutions, inverted bottlenecks, layer-scale, and GELU activations while preserving the strong locality bias of convolutions that CNNs naturally possess.

4. Experiments

4.1. Datasets and Protocol

We evaluate only the single-source cross-dataset protocol, where a model is trained on one source dataset and evaluated on a different target dataset. Multi-source protocols such as leave-one-out (LOO) and limited-source-domain (LSD) evaluation are not considered. We evaluate on four widely used MCIO datasets [54, 14, 42, 6] for intra-dataset ($M \rightarrow M$, $C \rightarrow C$, $I \rightarrow I$, $O \rightarrow O$) and cross-dataset ($M \rightarrow C$, $M \rightarrow I$, $M \rightarrow O$, ..., $O \rightarrow$ ...). MSU-MFSD[1] (**M**) includes videos of printed attacks, as well as replay attacks generated with a mobile phone and a tablet. CASIA-FASD[2] (**C**) includes warped-photo, cut-photo, and replay attacks under varying illumination. Replay-Attack[3] (**I**) provides high-resolution print-and-replay spoofs on a fixed display setup. OULU-NPU[4] (**O**) comprises six mobile phone sensors and multiple print-and-replay attack types, offering and the strongest sensor diversity.

Together, M/C/I/O cover laboratory and mobile capture, multiple PA, and the cross-dataset shift typical of PAD benchmarking [36, 18]. Per-split (train, dev, test) frame counts are in table 2 and Appendix section C.

Intra-dataset results are reported in tables 3 to 5 of this evaluation. cross-dataset results are reported in tables 4 and 5.

Pre-processing. We use MTCNN[5] [65] face detection preprocessing pipeline to crop faces in a size of 224×224 pixels from the video. Up to 20 frames are sampled evenly spaced across the video. Fig. 2 shows representative bonafide and attack crops from each MCIO corpus after this step. Each FM is paired with its image processor for resizing, normalization, and family-specific preprocessing. The most common input size is 224×224, although it varies across models: InternViT uses 448×448 inputs (Appendix B), while SigLIP-Large uses 256×256 inputs.

Models. Table 1 and table 3 list 24 frozen vision backbones. Only the linear probe is trained. Two end-to-end specialist PAD models (DeepPixBiS [23] and FSFM-FAS [51]) serve as specialist baselines under the same datasets and ACER metric (fully fine-tuned, not linearly probed).

Implementation details. Training minimizes `BCEWithLogitsLoss` on binary logits with class-balanced mini-batches (`WeightedRandomSampler`).

Table 2. MCIO frame counts (B/A = bonafide/attack).

	Train	Dev	Test
M	3196 (798/2398)	3037 (759/2278)	2398 (599/1799)
C	3600 (900/2700)	1200 (300/900)	7200 (1800/5400)
I	7199 (1200/5999)	7200 (1200/6000)	9600 (1600/8000)
O	24000 (4800/19200)	17999 (3600/14399)	12000 (2400/9600)

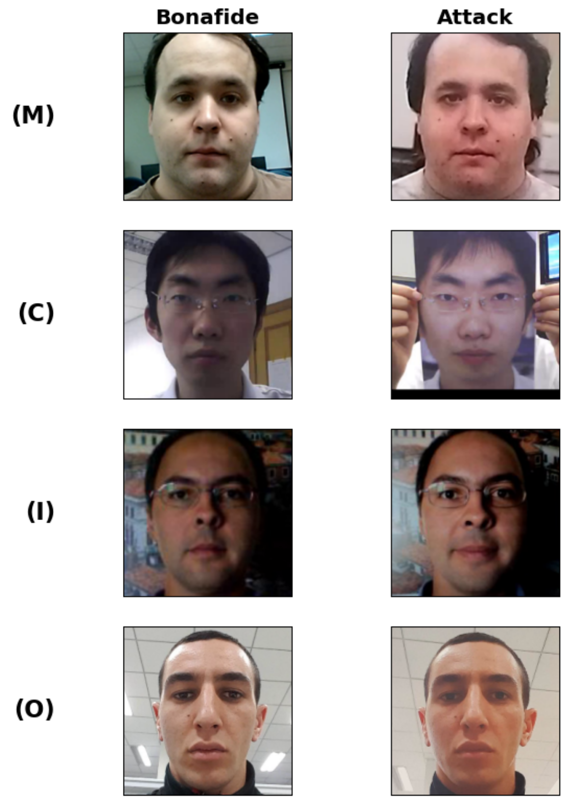


Figure 2. Examples face crops with the MTCNN detector [65]. Left column: bonafide. Right column: presentation attack (type: print). Rows - datasets: MSU-MFSD (M), CASIA-FASD (C), Replay-Attack (I), Oulu-NPU (O).

[1]M: Protocol `protocol` (all folds) shorturl.co/msu-mfsd-mod-2.2.9
[2]C: Protocol `grandtest` shorturl.co/casia-fasd-0b07ea45
[3]I: Protocol `grandtest` shorturl.co/replay-attack-aca6b46f
[4]O: Protocol `Protocol_1` shorturl.co/oulunpu-7bfb90c9
[5]MTCNN weights github.com/timesler/facenet-pytorch.

We optimize the probe with Adam (lr=10^{-4}, weight decay 10^{-6}), batch size 64, and a cosine learning-rate schedule (minimum 10^{-6}). Each run trains for up to 5–20 epochs with early stopping on validation loss (patience 8, minimum 5 epochs). Training-time augmentation applies random horizontal flip and color jitter (brightness, contrast, and saturation 0.15). We use PyTorch Lightning as the main training framework. The Bob toolkit [1, 2] provides the software library to calculate the EER and ACER metrics.

4.2. Evaluation metrics

We report the *Average Classification Error Rate* (ACER) following ISO/IEC 30107-3 [29] and standard PAD practice [36, 19, 18]. ACER is the mean of attack and bonafide error rates at a threshold τ tuned on the development split and applied unchanged to the test split:

$$\mathrm{ACER}(\tau) = \tfrac{1}{2}\big(\mathrm{APCER}(\tau) + \mathrm{BPCER}(\tau)\big). \quad (1)$$

Full APCER/BPCER definitions are in Appendix A. All reported ACER values are in percent (%) (lower is better). The operating threshold τ is selected on the development split by minimizing equal error rate (EER criterion), then applied to the test split. This dev-tuned thresholding is standard in PAD evaluation because it balances attack and bonafide error rates before reporting cross-corpus transfer.

Let $C = \{\mathrm{M}, \mathrm{C}, \mathrm{I}, \mathrm{O}\}$ denote the datasets and $A_{s,t}$ the test ACER (Eq. (1)) when the model is trained on corpus s and evaluated on corpus t. The full cross-dataset evaluation results appear in Appendix D (Tables 7–10). Table 4 shows two representative models in the main paper. We report two unweighted cross-dataset aggregates:

$$\mathrm{ACER}_{\mathrm{all}} = \frac{1}{|C|^2} \sum_{s \in C} \sum_{t \in C} A_{s,t}, \quad (2)$$

$$\mathrm{ACER}_{\mathrm{cross}} = \frac{1}{|C|(|C|-1)} \sum_{s \in C} \sum_{\substack{t \in C \\ t \neq s}} A_{s,t}. \quad (3)$$

Equation (2) averages all sixteen train/test pairs (including the intra-dataset diagonal). Equation (3) averages the twelve off-diagonal transfer pairs only. Table 5 reports Diag. (mean diagonal), Cross (mean off-diagonal), and Gap = Cross − Diag for each model.

4.3. Intra-dataset (ID) and Cross-dataset (CD) Performance

Intra-Dataset (ID). Table 3 reports test ACER in percent for intra-dataset training on MCIO datasets. The upper block lists frozen FMs in different sizes probed linearly. The lower block lists specialist baselines and a zero-shot MLLM InternVL-14B (containing InternViT-300M) models. A common trend is that larger foundation models (FMs) generally outperform smaller ones, which is also observed by Zhai et al. [62]. InternViT-6B achieves the lowest average intra-dataset ACER (1.6%). Among frozen foundation models, DINOv3-giant achieves the second-lowest average error (5.7%). However, the specialist baselines DeepPixBiS (3.4%) and FSFM-FAS (4.2%) still outperform most frozen probes. DeepPixBiS does so with only ~3M trainable weights, while InternViT has the largest linear probe with ~3k trainable weights. The InternVL-14B zero-shot model averages 28.8% ACER, far above even weak linear probes, indicating that generic VLM inference without a task-aligned readout is insufficient.

Cross-Dataset (CD). Full MCIO cross-dataset test ACER results for all benchmarked models are in Appendix D. Table 4 shows InternViT-6B and DeepPixBiS as representative cases, but CD (off-diagonal) entries remain high for nearly every model. When InternViT-6B is trained on MSU-MFSD, it reaches 4.4% on the diagonal but 14.7–45.3% on the three held-out corpora. DeepPixBiS trained on the same source shows a similar gap: 7.9% intra-dataset versus 36.5–49.7% off-diagonal. Even strong ID (diagonal cells) does not transfer reliably: DeepPixBiS trained on Replay-Attack reaches 0.0% intra-dataset yet 47.4–61.3% on other test corpora. InternViT-6B improves many pairs relative to InternViT-300M ($\mathrm{ACER}_{\mathrm{all}}$: 41.8% → 24.4%, $\mathrm{ACER}_{\mathrm{cross}}$: 47.7% → 32.0%) and follows the pattern that the larger model has better frozen features.

Gap. In Table 5, most models show moderate-to-high CD error despite varying ID performance. InternViT-6B achieves the lowest intra-dataset ACER (1.6% Diag.) but a 30.3% ACER gap, illustrating that strong linear separability within a corpus does not imply CD robustness. ViT-MSN occupies the opposite corner: near-chance intra-dataset ACER (>46%) with a small Gap (~9–10%), indicating uniformly poor representations rather than overfitting to a single corpus. DeepPixBiS and FSFM-FAS combine low ID error with large gaps (>33 points), indicating that fine-tuning can sharpen ID decision boundaries without improving CD generalization.

4.4. Accuracy vs compute trade-offs

Fig. 3 summarizes the accuracy–compute trade-off on the full MCIO cross-dataset evaluation. Larger backbones and higher-GFLOP encoders tend toward lower $\mathrm{ACER}_{\mathrm{all}}$. No single encoder dominates every budget. Mid-sized DINOv3 and CLIP variants offer practical compromises. CLIP ViT-B/32 is a notable outlier on this plot: with only ~90 M frozen parameters, it ranks second on $\mathrm{ACER}_{\mathrm{all}}$ (26.8%) and attains the lowest mean off-diagonal ACER in table 5 (31.6%), ahead of much larger CLIP and DINO variants. This pattern is plausible because contrastive image–text pretraining on diverse web-scale data encourages visual features that are less tied to a single acquisition pipeline [43], and frozen CLIP backbones have previously ranked among the strongest cross-dataset extractors in VFM-FAS benchmarks [47, 35, 22]. Within the

Table 3. Result: intra-dataset using the ACER (%) metric. The threshold τ (cf. eq. (1)) is calculated on the dev split and applied to the test split. The last two columns show the avg. is the unweighted mean and std. the standard deviation across the four MCIO corpora. The parameter sizes are marked with frozen ❄️ (in Billion - 10^9) for the foundation model and trainable 🔥 (in Kilo - 10^3) for the fully connected layer. The PAD specialist baselines (DeepPixBiS and FSFM-FAS) are fully trainable. The feature denotes how each frozen encoder is pooled to one vector (CLS token, pooling: post-layer-norm CLS, global average pooling, contrastive embedding, or mean-pooled patches) Datasets: M=MSU-MFSD, C=CASIA-FASD, I=Replay-Attack, O=OULU-NPU. Bold: lowest ACER per column.

Model	Feature	Parameters		Datasets					
		❄️ $\times 10^9$	🔥 $\times 10^3$	M	C	I	O	Avg.	Std.
Model Zoo									
ConvNeXt-V2-base [56]	GAP+LN	0.09	1.0	24.7	13.7	13.7	33.5	21.4	8.3
ConvNeXt-V2-huge [56]	GAP+LN	0.66	2.8	23.7	8.4	14.4	26.5	18.2	7.2
DINOv2-base [38]	CLS	0.09	0.8	20.2	2.7	9.8	20.2	13.2	7.4
DINOv2-giant [38]	CLS	1.14	1.5	10.8	5.5	3.6	13.8	8.4	4.1
DINOv2-with-registers-base [15]	CLS	0.09	0.8	16.9	7.6	4.9	28.9	14.6	9.4
DINOv2-with-registers-giant [15]	CLS	1.14	1.5	9.5	5.0	8.4	11.4	8.6	2.3
DINOv3-base [46]	CLS	0.09	0.8	12.8	3.8	6.1	14.9	9.4	4.6
DINOv3-giant [46]	CLS	0.84	1.3	8.8	**0.6**	7.9	5.5	5.7	3.2
BEiT-base [5]	Post-LN CLS	0.09	0.8	12.7	6.6	8.6	8.9	9.2	2.2
BEiT-large [5]	Post-LN CLS	0.30	1.0	14.5	9.0	5.0	7.9	9.1	3.4
ViT-MSN-base [3]	CLS	0.09	0.8	47.9	51.7	50.6	36.1	46.5	6.2
ViT-MSN-large [3]	CLS	0.30	1.0	49.7	51.0	42.7	37.5	45.2	5.5
ViT-MAE-base [26]	CLS	0.09	0.8	20.6	14.3	7.4	31.6	18.5	8.9
ViT-MAE-huge [26]	CLS	0.63	1.3	16.4	23.4	2.1	24.8	16.7	9.0
SigLIP-base [63]	Post-LN CLS	0.09	0.8	22.4	14.3	16.5	36.8	22.5	8.8
SigLIP-large [63]	Post-LN CLS	0.32	1.0	27.3	9.5	24.4	15.9	19.3	7.0
CLIP ViT-B/32 [43]	Post-LN CLS	0.09	0.8	21.8	4.4	18.6	5.7	12.6	7.7
CLIP ViT-L/14 [43]	Post-LN CLS	0.30	1.0	19.0	6.9	9.5	7.1	10.6	4.9
CLIP-ViT-H/14 [43]	Post-LN CLS	0.63	1.3	15.2	3.1	16.4	7.3	10.5	5.5
EVA-CLIP-8B [48]	Contrastive-embedding	8.22	1.3	12.2	2.1	15.3	11.2	10.2	4.9
Qwen2.5-VL-3B-ViT [4]	Mean pool	0.67	2.0	19.6	6.5	21.3	17.1	16.1	5.8
Qwen2.5-VL-7B-ViT [4]	Mean pool	0.68	3.6	17.2	6.1	27.6	18.1	17.2	7.6
InternViT-300M [12]	CLS	0.30	1.0	22.2	34.5	16.1	22.7	23.9	6.7
InternViT-6B [12]	CLS	5.54	3.2	4.4	1.8	0.1	**0.3**	**1.6**	**1.7**
Baseline									
InternVL3-14B (MLLM; 0-Shot) [12]	—	14.0	—	38.4	14.4	48.8	39.6	35.3	12.7
FSFM-FAS (MAE [26]; ViT-B16) [51]	—	—	86×10^3	**2.8**	1.6	0.2	12.0	4.2	4.6
DeepPixBiS (DNN) [23]	—	—	3×10^3	7.9	2.0	**0.0**	3.8	3.4	2.9

Table 4. Cross-domain ACER (%) for InternViT-6B and DeepPixBiS. The highest transferability is observed for InternViT-6B when trained on CASIA-FASD (C). Diagonal cells are intra-dataset (ID). Off-diagonal cells expose cross-dataset (CD) transfer. Full matrices for all benchmarked models are in Appendix D.

Model	Train	Test in ACER (%)					
		M	C	I	O	Avg.	Std.
InternViT-6B	M	4.42	16.64	45.29	14.65	20.25	15.18
	C	28.07	1.81	21.26	25.86	**19.25**	**10.36**
	I	28.54	25.26	0.06	50.33	26.05	17.83
	O	48.50	38.89	40.37	0.27	32.01	18.68
DeepPixBiS	M	7.89	47.05	49.65	36.46	35.26	16.56
	C	33.74	2.01	30.74	30.20	**24.17**	**12.87**
	I	47.36	61.34	0.00	50.94	39.91	23.61
	O	46.55	38.36	45.10	3.84	33.46	17.38

CLIP family, B/32 also outperforms ViT-L/14 and ViT-H/14 on cross-corpus transfer (31.6% vs. 34.0% and 38.7% mean off-diagonal ACER), suggesting that coarser 32×32 patch tokenisation can trade some intra-dataset separability for robustness to MCIO shift under a linear readout. The right panel of fig. 3 also highlights that CLIP ViT-B/32 and the compact CNN specialist DeepPixBiS [23] require almost the same forward-pass cost (~8.7 vs. ~9.2 GFLOPs at 224×224), yet frozen CLIP probing yields markedly lower cross-dataset error (ACER_{all} 26.8% vs. 33.2%; mean off-diagonal 31.6% vs. 43.1%). Similar inference budgets do not guarantee similar transfer: a task-trained specialist can match intra-dataset performance, but a frozen web-scale encoder offers stronger cross-corpus robustness at comparable compute.

4.5. Embedding Structure

Fig. 4 visualizes embedding structure for the two InternViT FMs. InternViT-6B forms more compact clusters than InternViT-300M, consistent with lower intra-dataset ACER. Yet, both retain strong dataset-specific structure that aligns with the cross-dataset error in table 5.

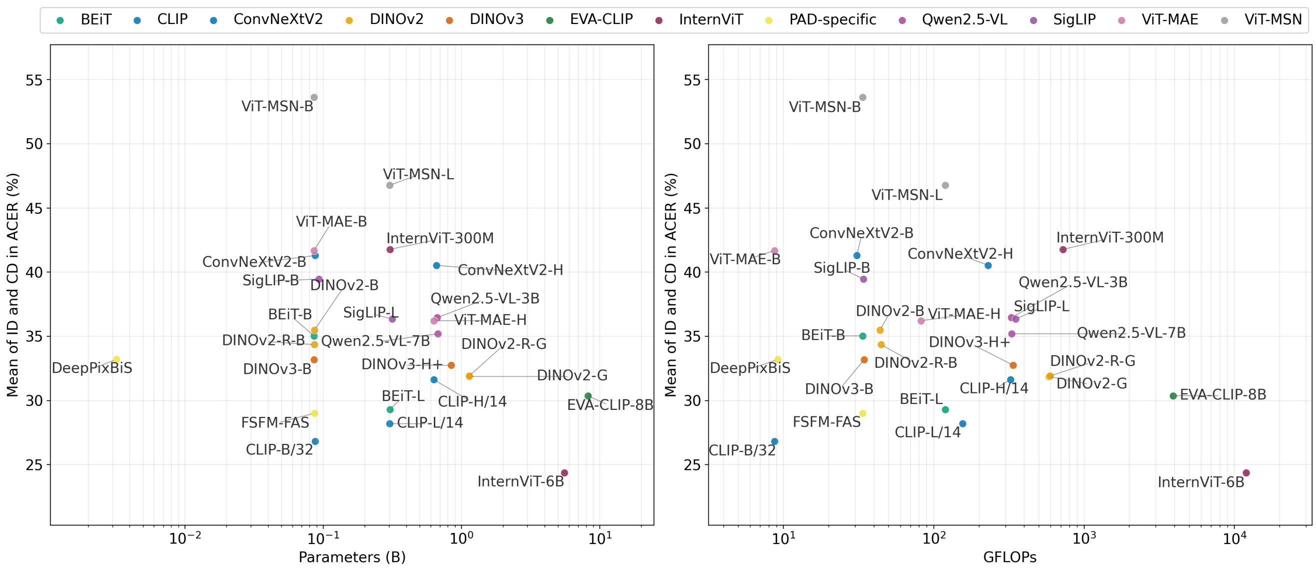


Figure 3. Comparison of the overall classification performance (intra-dataset and cross-dataset transfer measured in $ACER_{all}$ (%), Equation (2)) vs. GFLOPs per model. InternViT-6B has the second largest number of parameters and highest GFLOPs, while also achieving the best overall detection performance. Note that InternViT-6B processes images at approximately twice the resolution (see table 1), which is typically associated with a higher number of GFLOPs [43]. Although CLIP-B/32 has fewer GFLOPs than DeepPixBiS, its model size is much larger. CLIP-B/32 is also outperforming FSFM-FAS (backbone: CLIP-B/16), the specialized foundation model, on detection accuracy. Points are colored by model family. Lower ACER is better.

4.6. Discussion

Our results reveal a clear disconnect between intra-dataset performance and cross-dataset robustness. Models achieving near-perfect intra-dataset performance often exhibit substantial degradation under domain shift. InternViT-6B achieves the lowest error in intra-dataset settings, yet its mean cross-dataset ACER remains high (table 5), underscoring the difficulty of domain generalization even with massive backbones. FSFM-FAS, despite being pretrained specifically for face anti-spoofing, does not yield better PAD features than generic large-scale pretraining: it averages 4.2% intra-dataset ACER, higher than both InternViT-6B and DeepPixBiS. DeepPixBiS remains corpus-specific, strong on MSU-MFSD and Replay-Attack, weak on CASIA-FASD and OULU-NPU, while mid-sized frozen encoders such as DINOv3-giant and CLIP ViT-L/14 trade peak performance for lower variance across datasets (Table 3).

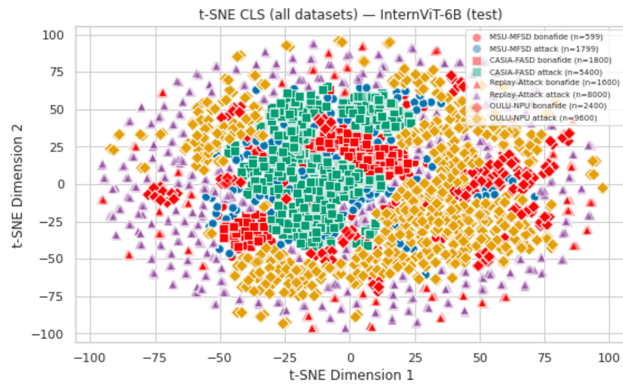
Pre-training objective appears at least as important as parameter count. Self-supervised and vision-language encoders (DINOv2/v3, BEiT, SigLIP, CLIP, EVA-CLIP) consistently outperform similarly sized ViT-MAE, ViT-MSN, and ConvNeXt-V2 variants on both intra- and cross-dataset ACER (Table 5). ViT-MSN performs near chance ($ACER_{all}$>53%) despite its ViT architecture, suggesting that not every large-scale pretraining objective yields PAD-relevant separable features. This aligns with the view that PAD benefits from encoders that preserve stable, localisable spatial structure. Feng *et al.* [22] argue that standard ViTs can misallocate attention to background regions, whereas DINOv2 with register tokens produces smoother feature maps better suited to fine-grained artefact detection [15]. Under our frozen linear-probe setting, the DINO and BEiT families already outperform reconstruction-based (ViT-MAE) and prototype-based (ViT-MSN) objectives, suggesting that self-distillation and discrete-token prediction preserve more PAD-relevant local structure.

Among families with multiple sizes (DINOv2, DINOv3, CLIP, SigLIP, InternViT), larger variants generally reduce $ACER_{all}$, though gains are family-dependent and non-monotonic. InternViT's superior performance may be partly attributed to its 448×448 pretraining resolution (Appendix B), yet this does not eliminate the cross-dataset gap (Table 5).

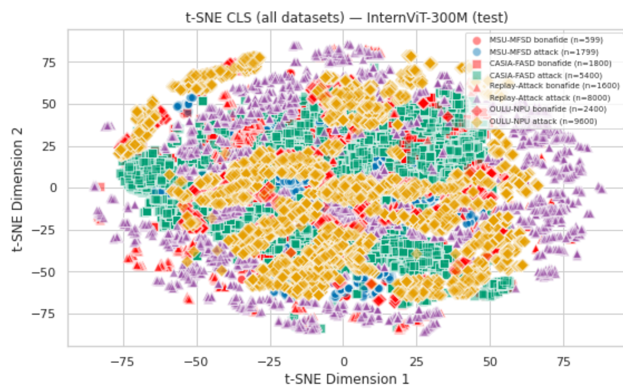
These frozen-probe scores represent a lower bound on what each backbone can achieve. Feng *et al.* [22] demonstrate that fine-tuning SSL backbones with spoof-specific augmentation (FAS-Aug) and patch supervision (PDA/APL) substantially reduces cross-dataset error on the same MCIO corpora. The large generalization gaps we report therefore reflect both limited source diversity in single-corpus training and the absence of task-specific adaptation, rather than an inherent limitation of pretrained vision models.

Table 5. Summary in the ACER (%) metric. Average over intra-dataset and cross-dataset (eq. (2)). Intra-dataset (ID) : mean of the four intra-dataset evaluation per model. Cross-dataset (CD) (eq. (3)): mean of the twelve cross-datasets (transferability). Gap = CD − ID. **Bold**: lowest value per column. All values in percent.

Model	All	ID	CD	Gap
InternViT-6B	**24.4**	**1.6**	32.0	30.3
CLIP ViT-B/32	26.8	12.6	**31.6**	18.9
CLIP ViT-L/14	28.2	10.6	34.0	23.4
FSFM-FAS	29.0	4.2	37.3	33.1
BEiT-large	29.3	9.1	36.0	26.9
EVA-CLIP-8B	30.4	10.2	37.1	26.9
CLIP ViT-H/14	31.6	10.5	38.7	28.2
DINOv2-giant	31.9	8.4	39.7	31.3
DINOv2-registers-giant	31.9	8.6	39.7	31.1
DINOv3-giant	32.7	5.7	41.8	36.1
DINOv3-base	33.2	9.4	41.1	31.7
DeepPixBiS	33.2	3.4	43.1	39.7
DINOv2-registers-base	34.4	14.6	40.9	26.4
BEiT-base	35.0	9.2	43.7	34.5
Qwen2.5-VL-7B-ViT	34.8	17.2	40.7	23.4
DINOv2-base	35.5	13.2	42.9	29.7
ViT-MAE-huge	36.2	16.7	42.7	26.1
SigLIP-large	36.3	19.3	42.0	22.8
Qwen2.5-VL-3B-ViT	36.5	16.1	43.2	27.1
SigLIP-base	39.5	22.5	45.1	22.6
ConvNeXt-V2-huge	40.5	18.2	48.0	29.7
ConvNeXt-V2-base	41.3	21.4	47.9	26.5
ViT-MAE-base	41.7	18.5	49.4	31.0
InternViT-300M	41.8	23.9	47.7	23.9
ViT-MSN-large	46.8	45.2	47.3	**2.0**
ViT-MSN-base	53.6	46.5	56.0	9.4



(a) InternViT-6B



(b) InternViT-300M

Figure 4. t-distributed stochastic neighbor embedding (t-SNE) of frozen embeddings (MCIO test, all corpora pooled). Bonafide: red. Attacks share the corpus hue. Umap plots in Appendix section F.

5. Conclusions

We benchmarked vision encoders of 24 frozen MLLMs, VFMs, and specialized networks for face PAD under a unified linear-probing protocol on the four MCIO datasets, reporting intra-dataset and cross-dataset evaluation with ACER (%), generalisation-gap analysis, and accuracy–compute trade-offs against PAD specialist baselines, i.e., the DNN DeepPixBiS and the ViT FSFM-FAS. Remarkably, several foundation-model encoders achieve competitive PAD performance in the intra-dataset setting without any backbone fine-tuning, demonstrating that spoof-related cues are already encoded in their pretrained representations. Despite recent studies on CLIP and DINO-based PAD systems, a broad and unified comparison across diverse foundation-model families and MLLM vision encoders remains limited [24, 22].

Our main findings are threefold. First, backbone scale and pretraining objective strongly affect intra-dataset detection accuracy. InternViT-6B reaches sub-percent error on several dataset splits with only one trainable FC layer. Second, cross-dataset transfer remains difficult, while intra-dataset error remains low for most FMs and DNNs. Third, model scale alone does not guarantee improved transferability. CLIP ViT-B/32 provides the strongest cross-dataset transfer–compute trade-off, achieving the lowest mean off-diagonal ACER among most evaluated probes while requiring substantially less computation. Compact specialists such as DeepPixBiS remain competitive at a fraction of the inference cost of multi-billion-parameter backbones, though they exhibit higher variance across source-target dataset pairs.

The results indicate that multimodal and face-spoofing-specialised pretraining do not automatically yield the most transferable linear features, while mid-sized SSL and vision–language encoders offer practical compromises when compute is limited. Future work includes intermediate-layer probing, multi-source protocols (LOO, LSD), and task-adaptive readouts beyond a single linear layer [40, 24, 22].

Acknowledgment

This research was funded by the European Union project CarMen (Grant Agreement No. 101168325).

References

[1] A. Andre, E. S. Laurant, W. Roy, G. Manuel, M. Chris, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *20th ACM Conference on Multimedia Systems (ACMMM)*, Oct. 2012.

[2] A. Andre, G. Manuel, d. F. P. Tiago, K. Pavel, M. Amir, and M. Sebastian. Continuously reproducing toolchains in pattern recognition and machine learning experiments. In *International Conference on Machine Learning (ICML)*, Aug. 2017.

[3] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, and N. Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pages 456–473. Springer, 2022.

[4] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[5] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[6] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pages 612–618. IEEE, 2017.

[7] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1578–1587, 2022.

[8] R. Cai, C.-Y. Soh, Z. Yu, H. Li, W. Yang, and A. C. Kot. Towards data-centric face anti-spoofing: Improving cross-domain generalization via physics-based data synthesis. *International Journal of Computer Vision*, pages 1–22, 2024.

[9] J. Cao and C. Ma. Towards generalized face anti-spoofing from a frequency shortcut view. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1005–1015. IEEE, 2025.

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[11] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[12] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[13] Z. Chen, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, F. Huang, and X. Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *AAAI*, pages 1132–1139. AAAI Press, 2021.

[14] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *2016 international conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2016.

[15] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski. Vision transformer needs registers. In *International Conference on Learning Representations*, 2024.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[18] M. Fang and N. Damer. Face presentation attack detection by excavating causal clues and adapting embedding statistics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6269–6279, 2024.

[19] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *IEEE/CVF WACV, Waikoloa, HI, USA, January 3-8, 2022*, pages 1131–1140. IEEE, 2022.

[20] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognit.*, 123:108398, 2022.

[21] M. Fang, M. Huber, and N. Damer. Synthaspoof: Developing face presentation attack detection based on privacy-friendly synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1061–1070, 2023.

[22] M. Feng, P. Gallin-Martel, K. Ito, and T. Aoki. Benchmarking vision foundation models for domain-generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1567–1576, 2026.

[23] A. George and S. Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 international conference on biometrics (ICB)*, pages 1–8. IEEE, 2019.

[24] L. J. Gonzalez-Soler, J. E. Tapia, and C. Busch. Are foundation models all you need for zero-shot face presentation attack detection? In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2025.

[25] X. Guo, Y. Liu, A. Jain, and X. Liu. Multi-domain learning for updating face anti-spoofing models. In *ECCV*, 2022.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[28] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In *European conference on computer vision*, pages 37–54. Springer, 2022.

[29] International Organization for Standardization. ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting, 2017.

[30] Y. Jia, J. Zhang, S. Shan, and X. Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, pages 8481–8490. Computer Vision Foundation / IEEE, 2020.

[31] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[32] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 13(7):1794–1809, 2018.

[33] Z. Li, T. Zhao, X. Xu, Z. Zhang, Z. Li, X. Chen, Q. Zhang, A. Bergamo, A. K. Jain, and Y. Xing. Optimal transport-guided source-free adaptation for face anti-spoofing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24351–24363, 2025.

[34] K.-H. Lin, Y.-W. Tseng, K.-Y. Huang, J.-C. Wu, and W.-H. Cheng. Instructflip: Exploring unified vision-language model for face anti-spoofing. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2987–2996, 2025.

[35] A. Liu, S. Xue, J. Gan, J. Wan, Y. Liang, J. Deng, S. Escalera, and Z. Lei. Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 222–232, 2024.

[36] Y. Liu, Y. Chen, W. Dai, C. Li, J. Zou, and H. Xiong. Causal intervention for generalizable face anti-spoofing. In *ICME*, pages 1–6. IEEE, 2022.

[37] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398. IEEE Computer Society, 2018.

[38] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[39] H. Otroshi Shahreza and S. Marcel. Foundation models and biometrics: A survey and outlook. *IEEE Transactions on Information Forensics and Security*, 2025.

[40] G. Ozgur, E. Caldeira, T. Chettaoui, F. Boutros, R. Ramachandra, and N. Damer. Foundpad: Foundation models reloaded for face presentation attack detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 745–755, 2025.

[41] D. Pasmino, C. Aravena, J. E. Tapia, and C. Busch. Flickr-pad: New face high-resolution presentation attack detection database. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2023.

[42] K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE transactions on information forensics and security*, 11(10):2268–2283, 2016.

[43] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[44] R. Shao, X. Lan, J. Li, and P. C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031. Computer Vision Foundation / IEEE, 2019.

[45] R. Shao, X. Lan, and P. C. Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, pages 11974–11981. AAAI Press, 2020.

[46] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[47] K. Srivatsan, M. Naseer, and K. Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19685–19696, 2023.

[48] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[49] C. Wang, Y. Lu, S. Yang, and S. Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In *CVPR*, pages 20249–20258. IEEE, 2022.

[50] G. Wang, H. Han, S. Shan, and X. Chen. Improving cross-database face presentation attack detection via adversarial domain adaptation. In *ICB*, pages 1–8. IEEE, 2019.

[51] G. Wang, F. Lin, T. Wu, Z. Liu, Z. Ba, and K. Ren. Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24364–24376, 2025.

[52] Z. Wang, Q. Wang, W. Deng, and G. Guo. Face anti-spoofing using transformers with relation-aware mechanism. *IEEE Trans. Biom. Behav. Identity Sci.*, 4(3):439–450, 2022.

[53] K. Watanabe, K. Ito, and T. Aoki. Spoofing attack detection in face recognition system using vision transformer with patch-wise data augmentation. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1561–1565, 2022.

[54] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.

[55] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[56] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.

[57] W. Yan, Y. Zeng, and H. Hu. Domain adversarial disentanglement network with cross-domain synthesis for generalized face anti-spoofing. *IEEE Trans. Circuits Syst. Video Technol.*, 32(10):7033–7046, 2022.

[58] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[59] J. Yang, X. Lin, Z. Yu, L. Zhang, X. Liu, H. Li, X. Yuan, and X. Cao. Dadm: Dual alignment of domain and modality for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12045–12056, 2025.

[60] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao. Revisiting pixel-wise supervision for face anti-spoofing. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(3):285–295, 2021.

[61] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao. NAS-FAS: static-dynamic central difference network search for face anti-spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(9):3005–3023, 2021.

[62] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[63] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[64] G. Zhang, K. Wang, H. Yue, A. Liu, G. Zhang, K. Yao, E. Ding, and J. Wang. Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[65] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[66] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European conference on computer vision*, pages 70–85. Springer, 2020.

[67] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A. ACER metric definitions

We follow ISO/IEC 30107-3 [29] and standard PAD practice [36, 19, 18]. Labels are bonafide (y=1) and attack (y=0). Let $\mathcal{P}_{\text{bon}}$ and $\mathcal{P}_{\text{att}}$ denote bonafide and attack test samples, and $\hat{y}_\tau(x) \in \{0, 1\}$ the hard decision at threshold τ. The *Attack Presentation Classification Error Rate* (APCER) and *Bonafide Presentation Classification Error Rate* (BPCER) are

$$\text{APCER}(\tau) = \frac{1}{|\mathcal{P}_{\text{att}}|} \sum_{x \in \mathcal{P}_{\text{att}}} \mathbb{I}[\hat{y}_\tau(x) = 1] , \quad (4)$$

$$\text{BPCER}(\tau) = \frac{1}{|\mathcal{P}_{\text{bon}}|} \sum_{x \in \mathcal{P}_{\text{bon}}} \mathbb{I}[\hat{y}_\tau(x) = 0] , \quad (5)$$

where $\mathbb{I}[\cdot]$ is the indicator function. ACER is $\text{ACER}(\tau) = \frac{1}{2}(\text{APCER}(\tau) + \text{BPCER}(\tau))$ (Eq. (1) in the main paper). The operating threshold is selected on the development split by minimizing $|\text{APCER}(\tau) - \text{BPCER}(\tau)|$ (EER criterion) and applied unchanged to the test split.

B. InternViT encoder details

InternViT is the vision tower of the InternVL open-source multimodal family [12, 11]. We use the standalone V2.5 checkpoints InternViT-300M and InternViT-6B, released without the language model or MLP connector (`InternVisionModel` in Hugging Face). Both are patch-based ViTs with 14×14 conv patch embedding, 448×448 input resolution, GELU feed-forward blocks, and pre-norm LayerNorm. InternViT-300M has 24 transformer layers, hidden size d=1024, 16 attention heads, and MLP width 4096 (~304 M parameters). InternViT-6B scales to 45 layers, d=3200, 25 heads, MLP width 12800, and QK normalization (~5.5 B parameters). We take the CLS token from the final layer as the global face embedding. The full InternVL stack (pixel unshuffle, dynamic tiling, MLP projector, and LLM) is not used at inference.

Pretraining follows the InternVL 2.5 recipe [11] in a ViT–MLP–LLM stack. In Stage 1, only the MLP connector is trained while the vision encoder and language model remain frozen. In optional Stage 1.5 (ViT incremental learning), the vision encoder and MLP are updated jointly using next-token prediction loss on multimodal data. In Stage 2 (instruction tuning), the full model is trained end-to-end on multimodal instruction datasets. The Hugging Face checkpoints retain weights from this joint vision–text pretraining. In our PAD experiments, the encoder is kept frozen, and only the linear probe is trained.

C. PAD dataset sample counts

Frame-level sample counts after face detection and cropping. We report the four MCIO datasets in fig. 6.

D. Cross-dataset generalizability

Full MCIO cross-dataset test ACER results for all benchmarked models. Each block lists one model; rows indicate the training corpus (M=MSU-MFSD, C=CASIA-FASD, I=Replay-Attack, O=Oulu-NPU) and columns the test corpus. Bold entries mark the lowest value in each column within the table (ACER in M–O; Avg./Std. as defined below; lower is better). Matrix aggregates ACER_{all} and $\text{ACER}_{\text{cross}}$ are defined in Eqs. (2)–(3) (summarized in Table 5). Avg. and Std. are the unweighted mean and population standard deviation across the four test corpora for each training source. All values are in percent (%); lower is better. Results can be found in the following tables 7 to 10.

E. GFLOPs measurement on DeepPixBis PAD detector on two image sizes (224 vs 448 pixels).

We compare two profiling libraries on three models in table 6: Torch[6] and Thop[7]. The difference between the GLOPs measurements of both libraries is small.

Table 6. GFLOPs comparison between Torch and THOP library. The number of GFLOPs increases four times by doubling the image size on DeepPixBis. The other models are rigid to the input size.

Model	Size	Torch GFLOPs	THOP GFLOPs	Diff (%)	Status
DeepPixBiS	224	9.169	9.304	1.47	ok
DeepPixBiS	448	36.676	37.215	1.47	ok
InternViT-6B	224	—	—	—	fail
InternViT-6B	448	11944.694	11339.534	-5.07	ok
CLIP-B/32	224	8.725	8.732	0.09	ok
CLIP-B/32	448	—	—	—	fail

F. CLS embedding t-SNE and UMAP visualizations

Each panel shows a two-dimensional t-SNE and UMAP projection of frozen CLS embeddings from the MCIO test split (MSU-MFSD, CASIA-FASD, Replay-Attack, Oulu-NPU), with all corpora pooled in one embedding space. t-SNE: Features are reduced with PCA (d=64) before t-SNE (perplexity 30, learning rate 200, 1000 iterations). UMAP: Features are reduced with PCA (d=64) before UMAP ($n_{\text{neighbors}}$=20, min_dist=0.1, cosine metric). Dataset-colored plots use distinct markers and hues per corpus; attack points within each corpus share the corpus color, and bonafide samples are shown in red.

Plot for InternVITs and CLIP-B/32 can be found in figs. 7 and 8.

[6] `from torch.utils.flop_counter import FlopCounterMode`

[7] https://github.com/ultralytics/thop

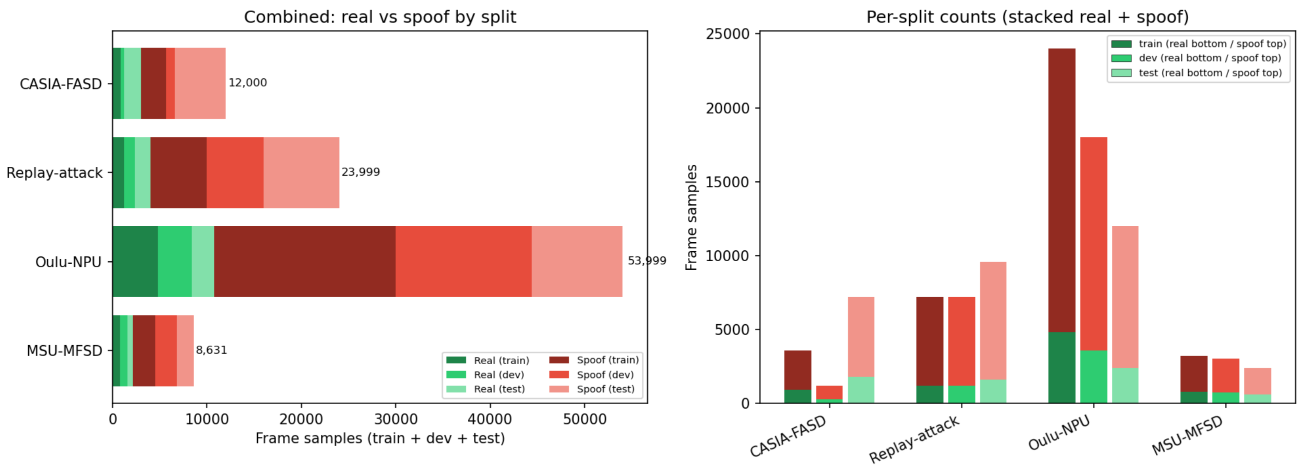


Figure 5. Combined summary

Figure 6. PAD dataset sample counts per protocol split (train, test, dev). Green: bonafide (real); red: attack (spoof). Counts are post-preprocess frame manifests (successful MTCNN extractions). Up to 20 frames per video.

Table 7. Cross-dataset test ACER (%) for MCIO leave-one-out training. Rows: training corpus; columns: test corpus. Bold: lowest ACER per column within the table. Avg./Std.: mean and population std. over the four test corpora per train source.

Model	Train	Test					
		M	C	I	O	Avg.	Std.
CLIP ViT-B/32	M	21.81	48.25	48.11	21.84	35.00	13.18
	C	27.35	4.39	34.77	18.68	21.30	11.30
	I	26.61	36.70	18.62	27.44	27.34	**6.41**
	O	25.69	24.02	39.30	**5.74**	23.69	11.94
CLIP ViT-L/14	M	18.99	44.13	35.04	43.86	35.50	10.21
	C	27.30	6.94	24.75	19.16	**19.54**	7.85
	I	29.21	35.45	9.54	46.28	30.12	13.36
	O	19.60	36.77	46.96	7.11	27.61	15.35
CLIP ViT-H/14	M	15.23	35.42	41.25	36.71	32.15	10.01
	C	30.52	3.09	33.98	41.34	27.23	14.47
	I	35.18	31.64	16.38	43.60	31.70	9.86
	O	47.66	39.80	46.97	7.29	35.43	16.54
DINOv2-base	M	20.16	36.82	46.41	37.62	35.25	9.49
	C	43.05	**2.67**	43.31	49.26	34.57	18.59
	I	35.78	32.99	9.76	44.19	30.68	12.76
	O	52.27	42.66	50.61	20.25	41.45	12.77
DINOv2-giant	M	10.84	32.76	45.68	33.38	30.67	12.55
	C	40.83	5.46	35.99	33.05	28.83	13.78
	I	45.33	25.59	**3.59**	47.40	30.48	17.70
	O	48.62	38.78	48.91	13.78	37.52	14.30
DINOv2-with-registers-base	M	16.90	27.07	38.44	39.69	30.53	9.28
	C	43.17	7.63	43.79	46.55	35.29	16.02
	I	38.76	31.82	4.86	50.51	31.49	16.76
	O	48.60	35.18	47.81	28.91	40.12	8.38
DINOv2-with-registers-giant	M	**9.51**	35.44	41.39	39.11	31.37	12.80
	C	27.08	5.02	39.52	35.02	26.66	13.26
	I	32.71	46.29	8.39	48.22	33.90	15.90
	O	44.37	39.89	47.15	11.40	35.70	14.27

Table 8. Cross-dataset test ACER (%) — continued.

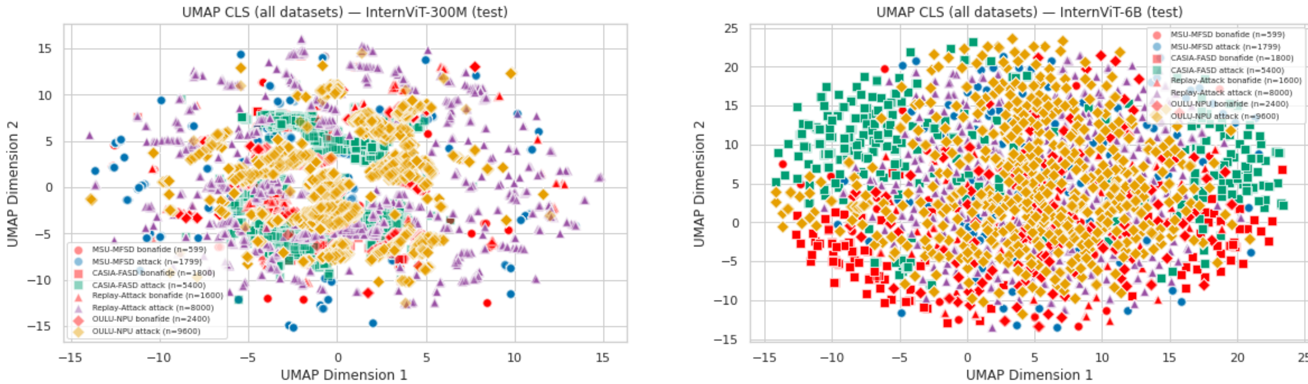
Model	Train	Test					
		M	C	I	O	Avg.	Std.
DINOv3-base	M	12.79	41.07	44.58	30.16	32.15	12.38
	C	44.89	3.81	38.42	38.41	31.38	16.13
	I	49.72	47.94	6.08	49.98	38.43	18.70
	O	30.43	39.59	38.17	14.89	30.77	9.81
DINOv3-giant	M	**8.76**	48.94	43.85	24.07	31.40	16.04
	C	40.61	**0.58**	45.17	33.09	29.86	17.45
	I	36.03	48.92	7.94	48.20	35.27	16.59
	O	35.56	46.71	50.00	**5.51**	34.44	17.54
BEiT-base	M	12.73	42.93	47.01	39.84	35.63	13.46
	C	47.22	6.57	39.12	49.82	35.68	17.26
	I	35.76	42.86	8.56	48.80	33.99	15.39
	O	47.10	35.64	47.79	8.91	34.86	15.74
BEiT-large	M	14.49	25.16	37.01	20.79	**24.36**	**8.23**
	C	44.27	9.05	38.96	18.23	27.63	14.48
	I	34.97	43.74	**5.02**	42.73	31.62	15.72
	O	43.54	36.51	46.59	7.87	33.63	15.31
ConvNeXt-V2-base	M	24.70	30.02	49.53	37.60	35.46	9.33
	C	44.18	13.68	56.89	47.34	40.52	16.19
	I	49.25	51.55	13.67	50.80	41.31	15.98
	O	47.81	51.47	58.68	33.47	47.86	9.18
ConvNeXt-V2-huge	M	23.66	46.81	40.49	35.34	36.58	8.49
	C	48.89	8.40	45.29	49.95	38.13	17.25
	I	52.55	54.17	14.37	50.27	42.84	16.50
	O	55.07	42.89	53.81	26.47	44.56	11.47

Table 9. Cross-dataset test ACER (%) — continued.

Model	Train	Test					
		M	C	I	O	Avg.	Std.
ViT-MSN-base	M	47.85	58.56	61.63	55.96	56.00	5.11
	C	54.09	51.70	64.04	58.04	56.97	**4.67**
	I	49.67	62.44	50.57	54.47	54.29	5.04
	O	52.04	52.57	48.18	36.06	47.21	6.66
ViT-MSN-large	M	49.75	52.44	46.80	39.01	47.00	5.03
	C	50.91	51.04	47.34	39.19	47.12	4.81
	I	50.17	53.39	42.66	39.58	46.45	5.56
	O	52.82	46.94	48.81	37.51	46.52	5.62
ViT-MAE-base	M	20.60	30.82	48.76	42.15	35.58	10.77
	C	47.86	14.32	64.76	48.09	43.76	18.32
	I	48.26	58.00	7.37	57.09	42.68	20.74
	O	54.44	37.09	55.73	31.59	44.71	10.56
ViT-MAE-huge	M	**16.38**	46.95	14.58	39.71	**29.40**	14.17
	C	47.53	23.40	41.72	47.15	39.95	9.83
	I	52.00	49.45	**2.07**	60.86	41.10	22.93
	O	38.02	39.85	34.98	24.81	34.41	5.81
SigLIP-base	M	22.36	39.90	46.04	38.36	36.67	8.75
	C	48.23	14.31	48.73	48.00	39.82	14.73
	I	43.05	38.94	16.46	51.70	37.54	13.01
	O	50.00	39.37	49.18	36.78	43.83	5.84
SigLIP-large	M	27.31	32.16	50.71	28.64	34.71	9.41
	C	48.39	**9.46**	47.98	48.65	38.62	16.84
	I	45.52	28.27	24.36	40.45	34.65	8.64
	O	47.52	37.30	48.88	**15.90**	37.40	13.20

Table 10. Cross-dataset test ACER (%) — continued.

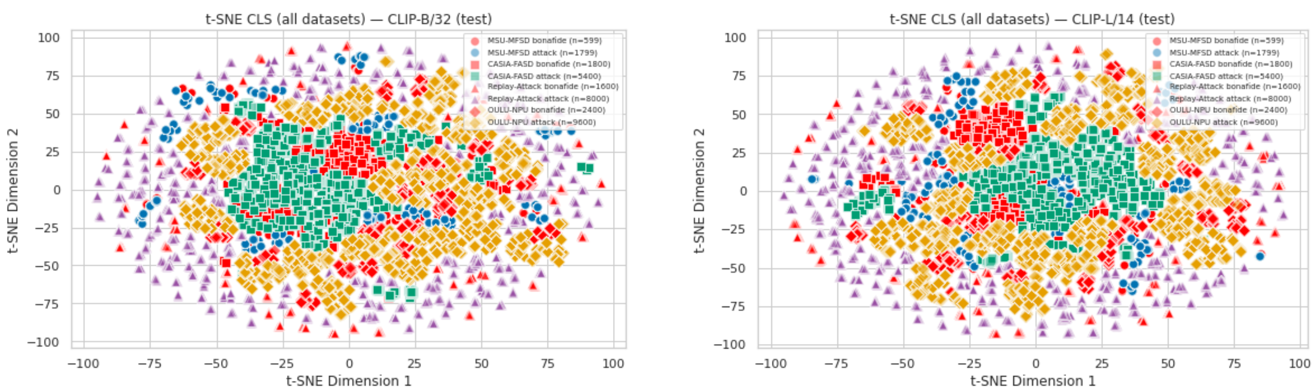
Model	Train	Test					
		M	C	I	O	Avg.	Std.
EVA-CLIP-8B	M	12.23	40.84	40.67	31.83	31.39	11.65
	C	30.02	2.13	25.90	39.92	24.49	13.88
	I	33.61	34.98	15.26	41.21	31.27	9.67
	O	46.21	31.00	49.00	11.15	34.34	15.04
Qwen2.5-VL-3B-ViT	M	19.61	35.64	50.17	33.45	34.72	10.83
	C	42.27	6.51	43.79	46.47	34.76	16.38
	I	53.02	44.80	21.33	48.65	41.95	12.25
	O	41.00	36.64	42.75	17.15	34.38	10.20
Qwen2.5-VL-7B-ViT	M	17.15	24.19	48.32	30.89	30.14	11.57
	C	43.76	6.08	46.48	38.19	33.63	16.18
	I	50.39	29.23	27.59	44.41	37.91	9.74
	O	50.91	38.47	42.70	18.14	37.55	12.07
InternViT-300M	M	22.17	40.42	51.18	47.18	40.24	11.12
	C	39.94	34.46	50.52	47.94	43.22	**6.39**
	I	50.06	50.83	16.07	50.19	41.79	14.85
	O	49.92	41.04	53.69	22.74	41.85	11.95
InternViT-6B	M	**4.42**	16.64	45.29	14.65	20.25	15.18
	C	28.07	**1.81**	21.26	25.86	**19.25**	10.36
	I	28.54	25.26	0.06	50.33	26.05	17.83
	O	48.50	38.89	40.37	**0.27**	32.01	18.68
FSFM_FAS	M	**2.84**	39.04	11.07	30.57	20.88	14.54
	C	43.81	**1.65**	52.28	38.97	34.18	19.38
	I	33.21	49.46	0.17	54.01	34.21	21.12
	O	39.28	21.97	33.76	12.04	26.76	10.55
DeepPixBiS	M	7.89	47.05	49.65	36.46	35.26	16.56
	C	33.74	2.01	30.74	30.20	24.17	12.87
	I	47.36	61.34	**0.00**	50.94	39.91	23.61
	O	46.55	38.36	45.10	3.84	33.46	17.38



(a) InternViT-300M

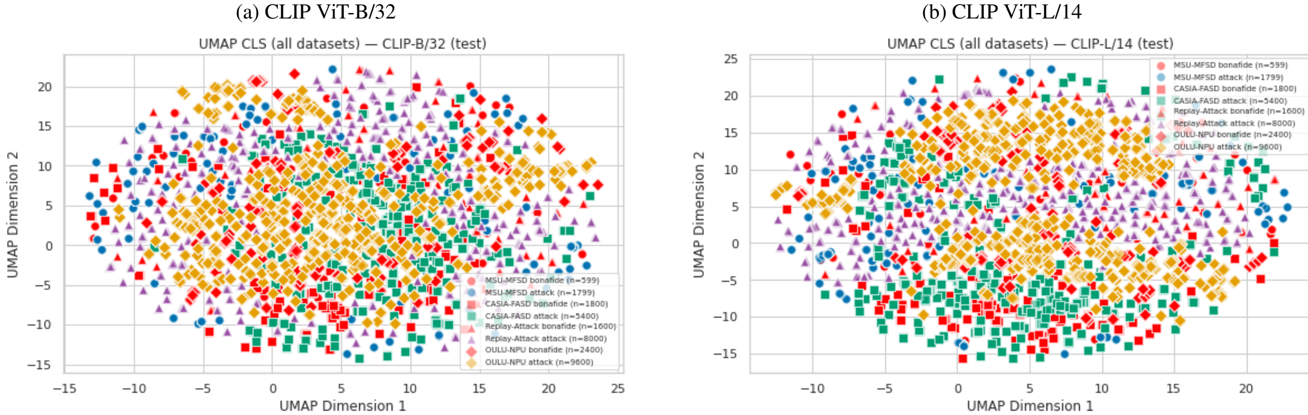
(b) InternViT-6B

Figure 7. t-SNE are in fig. 4. UMAP of frozen CLS embeddings (MCIO test samples, all corpora pooled; colored by dataset).



(a) CLIP ViT-B/32

(b) CLIP ViT-L/14



(c) CLIP ViT-B/32

(d) CLIP ViT-L/14

Figure 8. Top: t-SNE of frozen CLS embeddings (MCIO test samples, all corpora pooled; colored by dataset). Bottom: corresponding UMAP.