LoRA-based Adaptation Alone Is Not Enough: Understanding the Limits of Foundation Models for Face Presentation Attack Detection

Peter Lorenz[1], Anjith George[1], and Sébastien Marcel[1,2]

[1] Idiap Research Institute, Martigny, Switzerland
[2] University of Lausanne (UNIL), Lausanne, Switzerland

Abstract. Face presentation attack detection (PAD) aims to reliably detect a wide range of presentation attacks. While PAD methods achieve strong performance within individual datasets, their performance degrades under cross-dataset evaluation. Variations in sensors or lighting conditions can reduce the effectiveness of detectors from near-perfect to nearly random. Foundation models (FMs) have emerged as a promising alternative because typical PAD datasets, such as the MCIO benchmarks (MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU), are small relative to the scale used for web-based pretraining. However, existing PAD systems primarily focus on CLIP-based foundation models, while overlooking other FMs with different architectures and training procedures. This study addresses this question by systematically evaluating 32 FMs. Zero-shot prompting achieves performance near chance across model families and scales. The vision encoders, when low-rank-adapted (LoRA) with fewer than 1% trainable weights, achieve below 2% intra-dataset ACER in most cases, while cross-dataset ACER is substantially higher. LoRA primarily refines the decision boundary within a dataset, suggesting that the choice of encoder and adaptation dataset is more critical than adaptation alone. `https://gitlab.idiap.ch/pages/lora-pad`.

Keywords: LoRA · Presentation attack detection · Foundation models · Cross-dataset evaluation

1 Introduction

Face recognition has advanced rapidly with deep learning [7, 19], yet it remains vulnerable to presentation attacks: a printed photograph, a replayed video, or a similar spoof can induce a recognition system to grant access to an impostor [23, 50, 52, 81], or to reject a genuine user. Presentation attack detection (PAD) screens out spoofed presentations from genuine ones before verification [22, 47, 74, 75]. Detectors trained and tested on a single dataset achieve strong accuracy, but their performance drops sharply when the sensor, illumination, or attack type changes at test time [21, 35, 56, 57, 67, 72]. This cross-dataset gap, not intra-dataset accuracy, is the central open and long-lasting problem in PAD.

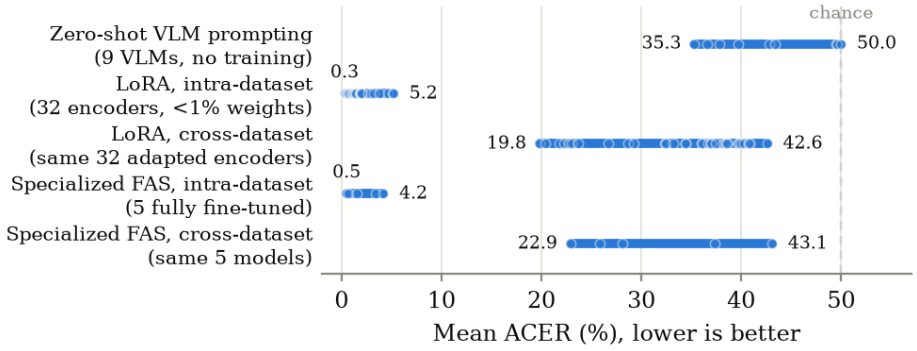


Fig. 1: Range of mean ACER across all evaluated models per setting (one dot per model, bar spans min–max). Zero-shot prompting of VLMs is near chance (Q1, Table 2). The same vision towers with LoRA (<1% trainable weights) reach 0.3–5.2% ACER intra-dataset (Q2, Table 3). Cross-dataset ACER remains at 19.8–42.6% across all 32 adapted encoders (Q3, Table 4). Our results suggest that LoRA alone cannot overcome dataset bias.

Vision encoders and vision–language models (VLMs) pretrained on web-scale image or image–text data [14, 37, 49, 54] encode substantially richer visual priors than the limited labelled data available in any PAD dataset [18, 21, 24], and recent work adapts them to PAD through parameter-efficient tuning such as LoRA [25, 29], multimodal architectures [44, 60, 78], and face-specific self-supervision [64] (Sec. 2). Whether such adaptation closes the cross-dataset gap or merely sharpens the decision boundary within the training dataset remains unresolved. As few encoders have been compared under a common parameter-efficient protocol [25, 29], the question stays open.

This benchmark includes 32 pretrained vision encoders, comprising vision-only models and the vision towers extracted from VLMs, each adapted with the same LoRA recipe on the four datasets, together with nine complete VLMs prompted zero-shot on the same splits and metric. Three specialist detectors serve as reference points from before the foundation-model era: DeepPixBiS [28], a classical CNN, FSFM-FAS [64], a face-pretrained ViT, and FLIP [60]. The goal is to characterize how representations behave across model families rather than to optimize dataset-specific accuracy.

Our contributions, framed by a single question: *is lightweight adaptation sufficient for robust PAD?*, are:

- *A comprehensive and systematic evaluation of foundation models for face PAD,* covering 32 vision encoders, 9 VLMs, specialist PAD baselines, and common datasets under a unified experimental protocol.
- *A bias free generalization score G:* We propose a metric G that jointly captures intra- and cross-dataset skill as the geometric mean below chance, which could be used as a future metric.

– *Unveiling the limitation of LoRA adaptation for PAD,* analyzing compute, feature-space analysis, and pretraining-scale comparisons besides accuracy.

2 Related Work

Face presentation attack detection (PAD) has evolved from developing task-specific representations using limited PAD datasets to leveraging large-scale pretrained models, thereby enhancing robustness under unseen conditions.

Learning representations from PAD data. Early PAD methods developed representations directly from PAD datasets using supervised objectives and appearance cues such as texture, color, and frequency information [47]. Subsequent approaches enhanced discrimination through pixel-wise supervision [28, 74], frequency-domain and multi-scale learning [22], and patch-based formulations [63]. While these methods are effective within the training distribution, their performance often declines in cross-dataset evaluations because they capture dataset-specific properties, such as illumination and attack artifacts, rather than learning transferable PAD representations [21, 67, 72].

To enhance generalization, subsequent research has explored self-supervised learning, domain adaptation, and domain generalization. Self-supervised methods reduce reliance on annotations by constructing supervisory signals from facial data, whereas domain adaptation leverages target-domain information during training [42, 65]. In contrast, domain generalization seeks to learn invariant representations from multiple source domains [15,35,56,57]. The use of auxiliary datasets and synthetic attack generation further increases data diversity [24,81]. Despite these efforts, multi-source evaluations consistently reveal substantial performance degradation on unseen datasets [25, 56], indicating that PAD-specific data alone may not yield sufficiently transferable representations.

Adapting large-scale pretrained representations for PAD. Recent research has examined whether large-scale pretrained representations offer greater transferability than models trained exclusively on PAD data. Vision foundation models pretrained through supervised, self-supervised, or multimodal objectives provide broad visual priors that can be adapted to PAD using various strategies [49,54]. Feng et al. [25] evaluated pretrained architectures under multi-source PAD protocols and demonstrated that self-supervised Vision Transformers generalize more effectively than supervised CNNs. Additional improvements have been achieved through architectural and training modifications, such as register tokens [17], PAD-specific augmentations, and patch supervision [9, 68].

In addition to full fine-tuning, parameter-efficient adaptation methods seek to retain pretrained knowledge by updating only a small subset of parameters. FoundPAD [29] adapts CLIP using LoRA [32], while vision-language methods such as FLIP [60], I-FAS [78], and InstructFLIP [44] incorporate language supervision. FaceCoT [79] explores multimodal large language models (MLLMs) by introducing a vision–language reasoning dataset and a chain-of-thought (CoT) enhanced learning strategy to improve spoof-detection robustness and interpretability. Other strategies include prompt learning [45], frequency-aware mod-

eling [10], domain-modality alignment [73], and optimal-transport adaptation [43]. FSFM-FAS [64] investigates face-specific masked modeling pretraining, while few-shot multimodal studies [38] indicate that semantic priors alone are insufficient for zero-shot PAD.

However, current foundation-model-based PAD studies assess only a limited range of backbone architectures, which complicates the separation of the effects of pretrained representations from those of adaptation strategies. FoundPAD and FLIP incorporate CLIP variants, while Feng et al. [25] compare a broader set of architectures but rely on full fine-tuning. As a result, it remains uncertain which pretraining paradigms, such as contrastive, self-supervised, or masked-image modeling, yield the most transferable representations for cross-dataset PAD and whether parameter-efficient adaptation can fully leverage these representations.

3 Approach

3.1 Low-rank Adaptation (LoRA) of the Vision Encoders

Vision encoders are evaluated using LoRA and a linear head, i.e., a fully connected layer (FC). Each cropped face image $\boldsymbol{x}$ is processed by a pretrained vision backbone with fixed weights during both training and testing, except for low-rank adapters applied to the attention projections. The backbone produces a global image embedding $\boldsymbol{h} \in \mathbb{R}^d$, with the extraction method determined by the specific backbone family. Extraction methods include CLS token, post-layer-norm CLS, global average pooling, projected CLIP image embedding, or mean-pooled patch tokens. Details are provided in the FC column of Tab. 3. The FC layer maps $\boldsymbol{h}$ to a scalar logit, $z = \boldsymbol{w}^\top \boldsymbol{h} + b$. Training minimizes binary cross-entropy loss on z, using bonafide labels y=1 and attack labels y=0. At inference, the bonafide probability is computed as $p = \sigma(z)$.

Following prior work [29], LoRA [32] reparameterizes each frozen projection $\boldsymbol{W}_0$ with a trainable low-rank update, $\boldsymbol{W}' = \boldsymbol{W}_0 + \frac{\alpha}{r}\boldsymbol{B}\boldsymbol{A}$, where rank r=8, scaling α=16, and dropout is set to 0.05. The matrix $\boldsymbol{B}$ is initialized to zero, ensuring adaptation begins from the pretrained behavior. Adapters are attached to the query and value projections ($\boldsymbol{W}_Q$, $\boldsymbol{W}_V$) of every attention block (`pwconv` layers for ConvNeXt; see the LoRA-site column in Tab. 3) and are trained jointly with the linear head. This approach maintains the trainable parameter footprint below 1% of the backbone weights while enabling the representation, rather than only the readout, to specialize for PAD.

3.2 Vision Encoders Family and Data Scale

Table 1 presents the 32 models, organized by family and scale, along with their pretraining provenance and data scale. Each backbone receives the same LoRA adaptation as described in Sec. 3.1. ViT-style backbones, including CLIP, DINOv2/v3, BEiT, SigLIP, ViT-MAE/MSN, and InternViT, utilize the final

CLS token. Vision towers extracted from multimodal large language models (Qwen-VL, LLaVA, DeepSeek-VL2) apply mean pooling over patch tokens, while ConvNeXt-V2 applies global average pooling.

The models encompass five pretraining regimes (Tab. 1): contrastive vision-language backbones (CLIP [54], SigLIP [77], EVA-CLIP [62]); self-distillation ViTs (DINOv2 [49], its register-token variant DINOv2-R [17], DINOv3 [58]); masked-image models (BEiT [5], ViT-MAE [31], ViT-MSN [3]); the modern convolutional neural network ConvNeXt-V2 [70]; and vision towers extracted from multimodal large language models (InternViT [14], Qwen-VL [4, 66], LLaVA-NeXT [46], LLaVA-OneVision [41], DeepSeek-VL2 [71]). These vision towers from the VLMs maintain parameter counts between 0.3B and 0.8B. For the DeepSeek-VL2 the vision encoder is also extracted. DeepSeek-VL2's vision encoder is a hybrid backbone that combines SigLIP-L and SAM-B [36]. SigLIP-L, trained with an image–text contrastive objective, captures high-level semantic concepts that align with language, and operates on a resolution of 384 pixels. SAM-B, trained for segmentation or mask prediction, captures low-level structure, edges, textures, and fine spatial details and operates on high resolution (1024×1024 pixels). Its outputs are mapped by a vision-language adaptor into a sequence of visual tokens that are then fed into the MoE large language model (LLM). DeepSeek-VL2-small and VL2-7B use the same SigLIP vision encoder. They differ only in the MoE language backbone size (2.8B vs 4.5B active parameters).

Three representative models are tracked throughout the experiments: **CLIP** (contrastive), **DINOv2-R** (self-supervised), and **LLaVA-NeXT** (vision-language model tower). More precisely, the multimodal LLaVA-NeXT v1.6 Vicuna with 7B parameters includes the pre-trained CLIP-ViT-Large (ViT-L/14) with a 336×336 pixel input resolution, whereas the other CLIP variants use 224×224 pixel input resolution. LLaVA-Next system trains in two stages: 1) Language-image alignment: The 2-layer ReLU MLP projects CLIP image features into the LLM's token space. 2) Visual instruction tuning: The full model (LLM, CLIP vision encoder, projector) is trained together but with different learning rates. CLIP is subsequently fine-tuned end-to-end with the LLM using a next-token prediction loss on multimodal instruction data, with a conservative learning rate to keep the LLM stable.

3.3 Specialist PAD Baselines

To contextualize the LoRA results, we compare them with four end-to-end specialist PAD models spanning classical CNN supervision and SSL-pretrained ViT fine-tuning: DeepPixBiS [28], a compact DNN with pixel-wise supervision. FSFM-FAS [64], a ViT-B/16 fine-tuned after face-specific SSL pretraining on the dataset VGGFace2 [11, 33]. FLIP [60] introduces three variants that differ in how much of the pretrained vision–language model is adapted: FLIP-V fine-tunes only the vision encoder. FLIP-IT fine-tunes both the image and text encoders. FLIP-MCL further fine-tunes the non-linear projection layer and both encoders, using language guidance to improve cross-dataset generalization. All

Table 1: Pretraining provenance of the vision backbones in Tab. 3. Variants listed together (e.g. base/large) share the same pretraining recipe and data. *Data scale* is the pretraining corpus size as reported by the underlying publication. VLM token budgets cover the full multimodal pretraining, of which the vision tower sees only the visual share. Params reports the frozen vision backbone size in billions (B), per listed variant. SSL: self-supervised learning, MIM: masked image modeling, NTP: next-token prediction, FT: fine-tuning.

Model	Params (B)	Descripton	Pretraining data	Data scale
CLIP ViT-B/32 or 16, ViT-L/14 [54]	0.09, 0.30	Contrastive image–text	WIT-400M [54]	400M pairs
CLIP ViT-H/14 [55]	0.63	Contrastive image–text	LAION-2B [55] (English subset of LAION-5B)	2.3B pairs, 32B seen
BEiT-base, -large [5]	0.09, 0.30	MIM (dVAE tokens), supervised FT	ImageNet-22K → ImageNet-1K [18]	14M images
ConvNeXt-V2-base, -huge [70]	0.09, 0.66	FCMAE (SSL), supervised FT	ImageNet-1K [18]	1.28M images
ViT-MAE-base, -huge [31]	0.09, 0.63	Masked autoencoding (SSL)	ImageNet-1K [18]	1.28M images
ViT-MSN-base, -large [3]	0.09, 0.30	Masked siamese networks (SSL)	ImageNet-1K [18]	1.28M images
DINOv2-base, -giant [49]	0.09, 1.14	DINO+iBOT SSL (base distilled from ViT-g)	LVD-142M [49] (curated web images)	142M images
DINOv2-R-base, -giant [17]	0.09, 1.14	DINOv2 SSL + register tokens	LVD-142M [49]	142M images
DINOv3-base, -huge+ [58]	0.09, 0.84	SSL + Gram anchoring (distilled from ViT-7B)	LVD-1689M [58] (curated Instagram/web pool)	1.69B images
SigLIP-base, -large [77]	0.09, 0.32	Sigmoid contrastive image–text	WebLI [12] (English subset)	∼10B images
EVA-CLIP-8B [62]	8.22	Contrastive image–text (EVA init)	Merged-2B [61] (LAION-2B [55] ∪ COYO-700M [8])	∼2B pairs
InternViT-300M [13]	0.30	Cosine distillation from InternViT-6B, NTP	InternVL 2.x multimodal corpus [13]	—
InternViT-6B [13]	5.54	CLIP contrastive, then incremental NTP	LAION-en/-multi [55], COYO [8], Wukong [30]; InternVL corpus [14]	∼5B pairs
Qwen2-VL-72B-ViT [66]	0.70	DFN ViT init, multimodal NTP	DFN-5B [20]; proprietary mix (OCR, interleaved, VQA, video)	1.4T tokens
Qwen2.5-VL-3B-, 7B-ViT [4]	0.67, 0.68	ViT from scratch (CLIP), multimodal NTP	DataComp [27] + in-house; proprietary multimodal mix	4.1T tokens
Qwen3-VL-2B-, 32B-ViT	0.41, 0.60	Multimodal NTP (no tech report)	Not disclosed	—
LLaVA-NeXT-7B-ViT [46, 54]	0.32	CLIP ViT-L/14-336 tower, lightly tuned in SFT	WIT-400M [54]; ∼1.3M instruction samples	400M pairs
LLaVA-OneVision-7B-ViT [41]	0.41	SigLIP-SO400M tower, unfrozen in later stages	WebLI [12]; ∼9.4M OneVision samples	—
DeepSeek-VL2-small-, 7B-ViT [71]	0.65, 0.78	SigLIP-SO400M init, VL pretraining	WebLI [12] + (WiT [59], WikiHow [39], OBELICS [40], in-house)	800B tokens

are fully fine-tuned on each dataset under the same protocol splits as our adapted backbones. FoundPAD [51] investigates LoRA-fine-tuned transformers, mainly limited to ViT-B/16 and ViT-L/14 on the leave-one-out protocol.

4 Experiments

4.1 Datasets and Protocol

This study evaluates only the single-source protocol, in which a model is trained on one dataset and tested either on the same dataset (intra-dataset) or on a different dataset (cross-dataset). Multi-source protocols, including leave-one-out and limited source domains, are not considered in this analysis. The four MCIO datasets [6, 16, 53, 69]: MSU-MFSD[3] (**M**), CASIA-FASD[4] (**C**), Replay-Attack[5] (**I**), and OULU-NPU[6] (**O**) span laboratory and mobile capture with print, replay, warped- and cut-photo, and multi-sensor attacks under varying illumination, giving the cross-dataset shift typical of PAD benchmarking [21, 48]. After face detection and cropping, the train/dev/test splits contain 3,196/3,037/2,398 (M), 3,600/1,200/7,200 (C), 7,199/7,200/9,600 (I), and 24,000/17,999/12,000 (O) frames, with attacks outnumbering bonafide frames by roughly 3–5× in every split.

Pre-processing. The MTCNN[7] [80] face detection preprocessing pipeline is employed to crop faces to a size of 224×224 pixels from the video frames. A

[3] M: Protocol `protocol` (all folds) `bob.db.msu_mfsd_mod-2.2.9.zip`
[4] C: Protocol `grandtest` `pad-face-casia-fasd.tar.gz`
[5] I: Protocol `grandtest` `pad-face-replay-attack.tar.gz`
[6] O: Protocol `Protocol_1` `pad-face-oulunpu.tar.gz`
[7] MTCNN weights github.com/timesler/facenet-pytorch.

maximum of 20 frames are sampled at evenly spaced intervals throughout the video. Each backbone model is paired with its corresponding image processor to perform resizing, normalization, and family-specific preprocessing. The most common input size is 224×224 pixels. However, this parameter varies across models. For example, InternViT uses 448×448 inputs, whereas SigLIP-Large uses 256×256 inputs.

Models. In the LoRA experiments (Tab. 3), the low-rank adapters and the linear head are trained. The five specialist PAD models are fully fine-tuned (Sec. 3.3) on our datasets.

Implementation details. We optimize the LoRA adapters and the linear head with Adam (binary cross-entropy, class-balanced mini-batches, batch size 32, lr=10^{-4} with cosine decay, early stopping on validation loss). The Bob toolkit [1, 2] computes the EER on the development set and ACER metrics on the test set.

4.2 Evaluation Metrics

The *Average Classification Error Rate* (ACER) is reported in accordance with ISO/IEC 30107-3 [34] and established PAD protocols [21,22,48]. ACER is defined as the mean of attack and bonafide error rates, calculated at a threshold τ determined on the development split using the Equal Error Rate (EER) criterion. This approach, standard in PAD evaluation, ensures balanced error rates and is applied without modification to the test split.

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

All ACER values are reported as percentages, where lower values indicate better performance.

Let $\mathcal{C} = \{\mathrm{M}, \mathrm{C}, \mathrm{I}, \mathrm{O}\}$ denote the datasets and $A_{s,t}$ the test ACER (Eq. (1)) of a model trained on dataset s and evaluated on dataset t. From the resulting transfer matrix $\boldsymbol{A} \in \mathbb{R}^{4\times4}$ we report three unweighted means:

$$\mathrm{ACER}_{\mathrm{ID}} = \tfrac{1}{4}\textstyle\sum_s A_{s,s}, \quad \mathrm{ACER}_{\mathrm{CD}} = \tfrac{1}{12}\sum_{s\neq t} A_{s,t}, \quad \mathrm{ACER}_{\mathrm{ALL}} = \tfrac{1}{16}\textstyle\sum_{s,t} A_{s,t}, \tag{2}$$

over the four intra-dataset (ID) pairs [6,28], the twelve cross-dataset (CD) transfer pairs [26,65,76], and all sixteen test evaluations (ID and CD combined). Each $A_{s,t}$ is computed on the test split of dataset t.

To summarize joint intra- and cross-dataset metrics in a single number, we report the *generalization score*

$$G \;=\; \frac{\sqrt{(C - \mathrm{ACER}_{\mathrm{ID}})_+ \,(C - \mathrm{ACER}_{\mathrm{CD}})_+}}{C}, \qquad C = 50, \;\; (x)_+ = \max(0, x), \tag{3}$$

the geometric mean of ID and CD *skill* (percentage points below the C=50% chance ACER), normalized to $[0, 1]$: G=1. Intuitively, G captures how far a model is from chance in both settings simultaneously. Since the product of two positive numbers is zero if either factor is zero, a model performing at chance on

Table 2: Zero-shot VLM PAD: intra-dataset test set ACER↓ (%). Prec. reports weight storage and compute dtype (e.g. NF4/`bf16`: 4-bit NF4 weights with `bfloat16` matmuls; `bf16`/`fp16`: full-precision loading). Avg. is the unweighted mean, and Std. is the standard deviation across the datasets.

Model	Prec.	M	C	I	O	Avg.	Std.
DeepSeek-VL2-small	NF4/fp16	43.8	22.4	46.9	45.7	39.7	10.1
DeepSeek-VL2	NF4/fp16	50.9	38.1	49.3	32.9	42.8	7.5
Qwen2.5-VL-7B	NF4/bf16	49.9	47.9	50.1	50.0	49.5	0.9
Qwen2.5-VL-3B	NF4/bf16	50.0	50.0	50.0	50.0	50.0	**0.0**
Qwen3-VL-2B	NF4/bf16	52.3	28.8	**42.0**	50.3	43.4	9.3
Qwen3-VL-32B	NF4/bf16	**37.0**	19.8	48.4	46.1	37.8	11.2
LLaVA-NeXT-7B	NF4/bf16	49.6	48.9	50.0	49.8	49.6	0.4
LLaVA-OneVision-7B	NF4/bf16	38.8	30.8	44.8	**32.1**	36.6	5.6
InternVL3-14B	NF4/bf16	38.4	**14.4**	48.8	39.6	**35.3**	12.7

ID will necessarily have $G = 0$, regardless of its CD ACER. The score rises only when skill is present on both fronts. Table 4 reports ID, CD, and G per model.

4.3 Zero-Shot VLM Prompting Performance

Table 2 presents results for prompting the VLM setting. No PAD-specific training is performed in this evaluation.

Prompt

```
"Analyze the following image and determine whether it is bonafide
or a presentation attack. Must provide an authenticity score
along with 'bonafide' or 'attacked' on a scale from 0 to 1, in
the format: score: xx, where xx is a number between 0 and 1. No
extra text."
```

In this configuration, full vision-language models (VLMs) are loaded using quantized NF4 precision and employ `bf16` or `fp16` computation. Each test image is evaluated with the fixed prompt (see above).

Are VLMs capable of solving PAD without task-specific training? Among the nine evaluated VLMs, which span four model families and range from 2 billion to 32 billion total parameters (Tab. 2), the average ACER remains between 35% and 50%, which is close to chance performance. *The answer is no*: language-aligned semantic reasoning does not substitute for task-aligned adaptation of the vision tower, even though it is the very same tower.

4.4 Intra-dataset Performance under LoRA

Table 3 extends this recipe to 32 vision backbones (Tab. 1), from CLIP and DINOv3 to the vision towers of multimodal LLMs. The intra-dataset gains are

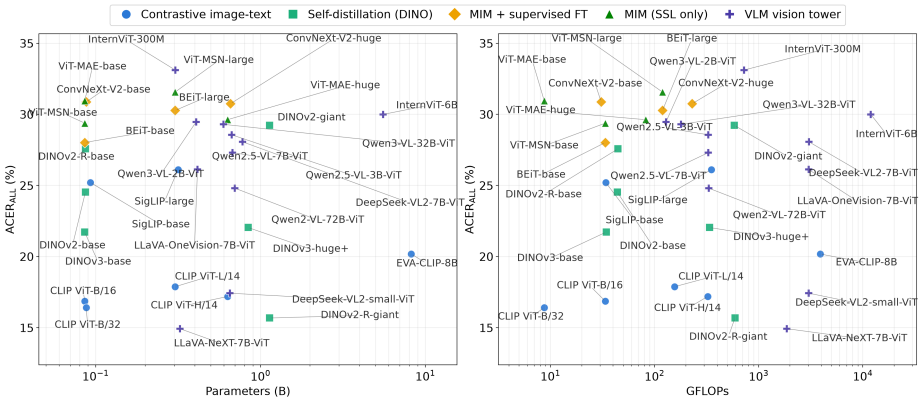


Fig. 2: ACER$_{\text{ALL}}$ ↓ (%) vs. compute↓ (GFLOPs). Left: frozen vision backbone parameters (billions). Right: forward-pass GFLOPs at batch size 1, where CLIP VIT-B/32 yields the best trade-off between accuracy and compute. Note that some jitter was added to the x-axis of the datapoints such that they do not overlap.

large and consistent: CLIP ViT-B/32 reaches 0.3% mean ACER, LLaVA-NeXT-7B-ViT 0.4%, InternViT-6B 0.7%, and most adapted backbones remain below 2% across the four datasets while updating fewer than 1% of the foundation model parameters. These results indicate that parameter-efficient adaptation is sufficient for strong intra-dataset PAD performance, although, as discussed in Sec. 4.5, the choice of backbone remains an important factor.

4.5 Which Backbones Benefit most from LoRA?

Figure 2 relates post-LoRA ACER to backbone size and forward-pass cost, addressing the second half of Q2. The pretraining objective matters more than parameter count: CLIP ViT-B/32, with a 90M-parameter backbone, reaches 0.3% mean intra-dataset ACER, whereas the 0.60B-parameter vision tower extracted from the Qwen3-VL-32B VLM averages 5.0% (Tab. 3). Contrastive vision–language backbones (CLIP, EVA-CLIP, SigLIP) and self-distillation models (DINOv3) respond most reliably to LoRA. The vision towers extracted from multimodal LLMs show mixed results across scales. Gains also vary by dataset: the weakest backbones (ViT-MAE, ViT-MSN, InternViT-300M) remain the weakest on OULU-NPU, the dataset with the strongest sensor diversity. Backbone choice and pretraining objective therefore dominate LoRA outcomes. Scale alone does not.

4.6 Cross-Dataset Transfer under LoRA

Does lightweight adaptation solve cross-dataset shift? While intra-dataset ACER demonstrates strong performance following LoRA adaptation, cross-dataset ACER remains between 20% and 43% for all models, irrespective of pretraining regime,

Table 3: ID test ACER↓ (%) with LoRA fine-tuning: each cell trains and tests on the same dataset (threshold tuned on the `dev` split, EER criterion). Features: adapter site (LoRA) and backbone feature fed to the fully connected head (FC). Parameters: frozen backbone size (❄, $\times 10^9$) and trainable LoRA/FC parameter counts (🔥, $\times 10^3$). Avg./Std.: unweighted mean and standard deviation across the datasets. **Bold**: lowest ACER per column.

Model	Features		Parameters			Test Datasets					
	LoRA	FC	❄$_{\times 10^9}$	🔥$_{(\text{LoRA, FC})\times 10^3}$		M	C	I	O	Avg.	Std.
Model Zoo - LoRA Finetuned											
ViT-MSN-large	Q,V	CLS	0.30	786	1.0	4.2	3.2	**0.0**	13.2	5.1	4.9
Qwen3-VL-32B-ViT	Q,V	Mean pool	0.60	995	5.1	9.2	0.5	5.3	4.9	5.0	3.1
InternViT-300M	Q,V	CLS	0.30	786	1.0	3.5	0.9	1.3	11.3	4.2	4.2
ViT-MAE-base	Q,V	CLS	0.09	294	0.8	1.4	1.6	0.4	10.9	3.6	4.3
Qwen2.5-VL-3B-ViT	Q,V	Mean pool	0.67	1310	2.0	3.8	0.5	1.2	6.9	3.1	2.5
ViT-MSN-base	Q,V	CLS	0.09	294	0.8	0.8	1.0	0.1	9.7	2.9	3.9
ViT-MAE-huge	Q,V	CLS	0.63	1310	1.3	2.8	0.6	0.1	5.4	2.2	2.1
Qwen2-VL-72B-ViT	Q,V	Mean pool	0.70	1310	8.2	1.8	1.2	4.4	0.8	2.1	1.4
DeepSeek-VL2-7B-ViT	Q,V	Mean pool	0.78	995	2.6	4.3	0.1	3.4	0.5	2.1	1.8
DINOv2-R-base	Q,V	CLS	0.09	294	0.8	0.6	2.1	0.6	4.7	2.0	1.7
BEiT-large	Q,V	LN CLS	0.30	786	1.0	0.6	0.6	0.5	6.1	1.9	2.4
Qwen3-VL-2B-ViT	Q,V	Mean pool	0.41	786	2.0	2.5	0.3	2.9	1.7	1.9	1.0
DINOv2-giant	Q,V	CLS	1.14	1966	1.5	2.3	0.5	0.5	3.9	1.8	1.4
Qwen2.5-VL-7B-ViT	Q,V	Mean pool	0.68	1310	3.6	2.9	0.2	2.3	0.3	1.4	1.2
SigLIP-large	Q,V	LN CLS	0.32	786	1.0	2.7	0.7	0.9	1.2	1.4	0.8
DINOv2-R-giant	Q,V	CLS	1.14	1966	1.5	2.9	0.2	1.7	0.8	1.4	1.0
DINOv2-base	Q,V	CLS	0.09	294	0.8	1.5	1.6	0.3	2.1	1.4	0.6
SigLIP-base	Q,V	LN CLS	0.09	294	0.8	0.9	0.7	2.0	1.2	1.2	0.5
LLaVA-OneVision-7B-ViT	Q,V	Mean pool	0.41	958.5	3.6	1.8	0.4	0.7	1.8	1.2	0.6
BEiT-base	Q,V	LN CLS	0.09	294	0.8	0.3	0.6	0.4	3.2	1.1	1.2
ConvNeXt-V2-base	pwconv	GAP+LN	0.09	1443	1.0	0.4	0.8	1.4	1.7	1.1	0.5
ConvNeXt-V2-huge	pwconv	GAP+LN	0.66	3970	2.8	1.8	0.5	0.3	0.9	0.9	0.6
DeepSeek-VL2-small-ViT	Q,V	Mean pool	0.65	995	2.0	1.7	0.1	**0.0**	1.1	0.7	0.7
EVA-CLIP-8B	Q,V	Img emb.	8.22	5505	1.3	0.6	0.5	0.8	0.9	0.7	0.2
CLIP ViT-H/14	Q,V	LN CLS	0.63	1310	1.3	0.3	**0.0**	2.3	0.2	0.7	0.9
DINOv3-base	Q,V	CLS	0.09	294	0.8	0.6	0.5	1.2	0.5	0.7	0.3
InternViT-6B	Q,V	CLS	5.54	4608	3.2	0.6	0.1	1.5	0.4	0.7	0.5
DINOv3-huge+	Q,V	CLS	0.84	1310	1.3	**0.0**	0.0	0.7	1.0	0.4	0.4
CLIP ViT-B/16	Q,V	LN CLS	0.09	294	0.8	0.4	1.1	0.2	**0.0**	0.4	0.4
LLaVA-NeXT-7B-ViT	Q,V	Mean pool	0.32	786	4.1	1.2	**0.0**	0.2	0.0	0.4	0.5
CLIP ViT-L/14	Q,V	LN CLS	0.30	786	1.0	**0.0**	0.1	**0.0**	1.4	0.4	0.6
CLIP ViT-B/32	Q,V	LN CLS	0.09	294	0.8	0.8	0.2	0.2	**0.0**	**0.3**	0.3
Baseline – PAD Specialist Models other than LoRA											
FSFM-FAS	—	—	—	86		2.8	1.6	0.2	12.0	4.2	4.6
DeepPixBiS	—	—	—	3		7.9	2.0	**0.0**	3.8	3.4	2.9
FLIP-V	—	—	—	86		**0.0**	**0.4**	0.8	4.7	1.5	1.9
FLIP-MCL	—	—	—	169		1.6	0.7	0.3	0.1	0.7	0.6
FLIP-IT	—	—	—	149		0.4	0.8	0.8	**0.1**	**0.5**	**0.3**

Table 4: ACER (%) over the datasets (Eq. (2)): ID (four intra-dataset pairs), CD (twelve transfer pairs), All (sixteen pairs; ID + CD), and generalization score G (higher is better). The geometric mean of intra-dataset and cross-dataset skill relative to the 50% chance ACER. FLIP (IT/MCL/V), FSFM-FAS, and DeepPixBiS are the fully fine-tuned FAS-specialist baselines (Sec. 3.3). **Bold**: best per column (lowest ACER, highest G).

Model	All	ID	CD	$G \uparrow$
LLaVA-NeXT-7B-ViT	**14.9**	0.4	**19.8**	**0.77**
DINOv2-R-giant	15.7	1.4	20.4	0.76
CLIP ViT-B/32	16.4	**0.3**	21.8	0.75
CLIP ViT-B/16 (FoundPAD)	16.9	0.5	22.3	0.74
CLIP ViT-H/14	17.2	0.7	22.7	0.73
DeepSeek-VL2-small-ViT	17.4	0.7	23.0	0.73
CLIP ViT-L/14 (FoundPAD)	17.9	0.4	23.7	0.72
EVA-CLIP-8B	20.2	0.7	26.7	0.68
DINOv3-base	21.7	0.7	28.7	0.65
DINOv3-huge+	22.0	0.4	29.2	0.64
DINOv2-base	24.5	1.4	32.3	0.59
Qwen2-VL-72B-ViT	24.8	2.0	32.4	0.58
SigLIP-base	25.2	1.2	33.2	0.57
SigLIP-large	26.1	1.4	34.4	0.55
LLaVA-OneVision-7B-ViT	26.1	1.2	34.4	0.55
Qwen2.5-VL-7B-ViT	27.3	1.2	36.0	0.52
DINOv2-R-base	27.6	1.9	36.1	0.52
BEiT-base	28.0	1.1	37.0	0.50
DeepSeek-VL2-7B-ViT	28.1	2.0	36.7	0.50
Qwen2.5-VL-3B-ViT	28.6	3.2	37.0	0.49
DINOv2-giant	29.2	1.8	38.4	0.47
Qwen3-VL-32B-ViT	29.3	5.1	37.4	0.48
ViT-MSN-base	29.4	2.9	38.2	0.47
Qwen3-VL-2B-ViT	29.5	1.9	38.6	0.47
ViT-MAE-huge	29.6	2.7	38.6	0.46
InternViT-6B	30.0	0.7	39.8	0.45
BEiT-large	30.3	1.9	39.7	0.44
ConvNeXt-V2-huge	30.7	0.9	40.7	0.43
ConvNeXt-V2-base	30.9	1.1	40.8	0.42
ViT-MAE-base	30.9	3.8	40.0	0.43
ViT-MSN-large	31.6	5.2	40.3	0.42
InternViT-300M	33.1	4.6	42.6	0.37
PAD Specialist Baseline other than LoRA				
FLIP-MCL	**17.3**	0.7	**22.9**	**0.73**
FLIP-IT	19.5	**0.5**	25.8	0.69
FLIP-V	21.5	1.5	28.1	0.65
FSFM-FAS	29.0	4.2	37.3	0.48
DeepPixBiS	33.2	3.4	43.1	0.36

parameter count, or intra-dataset rank. Transfer performance does not correlate with intra-dataset rank. LLaVA-NeXT-7B-ViT achieves the lowest mean cross-dataset ACER (19.8%), followed by DINOv2-R-giant (20.4%) and CLIP ViT-B/32 (21.8%). In contrast, InternViT-6B, which is among the strongest intra-dataset backbones, ranks near the bottom (39.8%). The choice of training source significantly influences transfer performance. For example, CLIP ViT-B/32 adapted on Replay-Attack achieves approximately 0.2% intra-dataset ACER and approximately 10% mean cross-dataset ACER, whereas the same model adapted on MSU-MFSD yields approximately 33% cross-dataset ACER. *Collectively, these findings indicate that lightweight adaptation does not resolve cross-dataset transfer.* Instead, the primary limitation shifts from representation quality to the selection of the backbone and the training dataset for the deployment domain.

4.7 Pretraining Data Scale

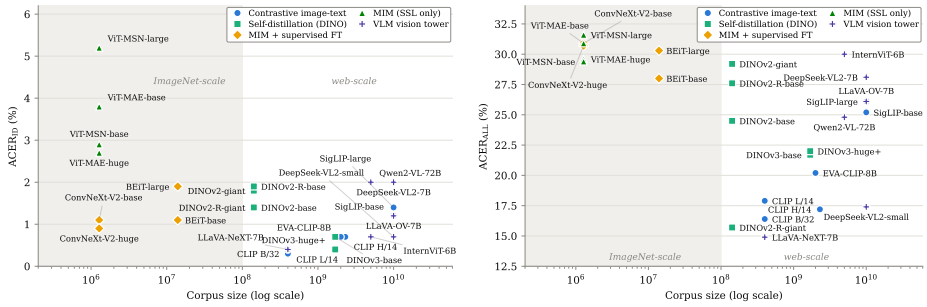
Figure 3 relates each backbone's pretraining dataset size to its post-LoRA error. The intra-dataset view (left) is the control for the data-scale argument: after LoRA, every backbone fits its training dataset (0.3–5.2% ACER), ImageNet-only backbones sit among the best intra-dataset models, and no left–right structure is visible, so pretraining dataset size is uninformative for intra-dataset performance. The all-pairs aggregate (right) reverses the picture: every backbone pretrained solely on ImageNet-1K/22K ($\leq$14M images, shaded) sits in the bottom third, every backbone in the top ten saw at least 100M web images, and beyond web scale more data stops helping (SigLIP's $\sim$10B WebLI images rank behind CLIP's 400M pairs). Pretraining data scale therefore manifests almost exclusively under domain shift: web-scale coverage is necessary but not sufficient, and curation and objective dominate beyond the threshold.

4.8 Embedding Structure

Figure 4 presents the adapted embeddings of the three most effective backbone models (CLIP ViT-B/32, DINOv2-R-giant, LLaVA-NeXT-7B-ViT) across the test sets. Although the attention projections are adapted, the feature geometry is still primarily structured by acquisition domain rather than by presentation-attack class. Samples cluster according to their source dataset rather than the bonafide or attack label. This representational structure accounts for the persistent cross-dataset performance degradation observed in Tab. 4. While the low-rank update enhances separation within individual datasets, it does not create a unified spoof-versus-live representation across datasets. As a result, thresholds calibrated in one domain are unlikely to generalize effectively to another.

4.9 Discussion

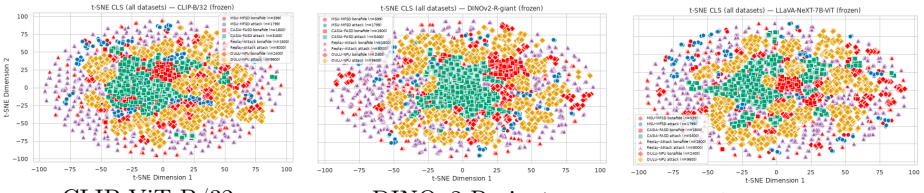
Why does adaptation saturate intra-dataset but not cross-dataset? Zero-shot VLM prompting does not achieve viable performance (Tab. 2), while



(a) $\text{ACER}_{\text{ID}} \downarrow$ **(b)** $\text{ACER}_{\text{ALL}} \downarrow$

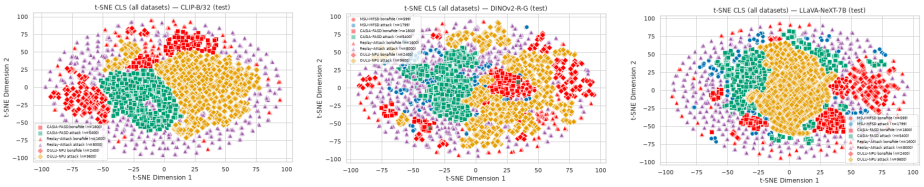
Fig. 3: Relationship between dataset size and ACER. The mean intra-dataset error (ACER_{ID}, left) and cross-dataset error (ACER_{ALL}, right) are presented. Colors and markers indicate pretraining families (see Sec. 3.2). The shaded band denotes ImageNet-scale datasets ($\leq 10^8$ images). Intra-dataset performance is insensitive to dataset size, while cross-dataset performance plateaus beyond 10^8 images.

(a) Off-the-shelf (frozen backbone)



CLIP ViT-B/32 DINOv2-R-giant LLaVA-NeXT-7B-ViT

(b) After LoRA fine-tuning



CLIP ViT-B/32 DINOv2-R-giant LLaVA-NeXT-7B-ViT

Fig. 4: t-SNE of CLS features on the test sets. **(a)** Off-the-shelf frozen backbones (before adaptation); **(b)** the same encoders after LoRA fine-tuning. Bonafide samples are shown in red across all datasets, while presentation attacks are colored: MSU-MFSD, CASIA-FASD, Replay-Attack, and OULU-NPU. Residual domain structure behind the cross-dataset gap.

LoRA adaptation enables strong discrimination within individual acquisition pipelines (Tab. 3). LoRA primarily enhances intra-dataset discrimination, whereas cross-dataset invariance requires additional mechanisms. Nevertheless, LoRA adaptation does not eliminate acquisition-specific cues that dominate the learned representations. The resulting embeddings are still organized by the source dataset rather than by the spoof label (Fig. 4), leading to thresholds calibrated on one dataset that fail under distribution shift. As a result, the cross-dataset ACER remains between 20 and 43 percent across all model backbones (Tab. 4), despite very low intra-dataset error rates. For example, InternViT-6B achieves 0.7 percent intra-dataset ACER but 39.8 percent cross-dataset ACER.

Why do scale and pretraining volume not rescue transfer? Cross-dataset transfer is determined more by the pretraining objective than by model size or pretraining scale. Contrastive vision-language encoders (CLIP, EVA-CLIP) and the DINO family, particularly the register-token variant DINOv2-R-giant, consistently outperform similarly sized ViT-MAE, ViT-MSN, and ConvNeXt-V2 models (Tab. 4). These findings align with evidence that PAD benefits from encoders preserving stable, localizable spatial structure [17,25]. Increasing model size provides little benefit: the 0.09B CLIP ViT-B/32 surpasses every multi-billion-parameter backbone, and smaller variants often transfer better within the same family. Likewise, increasing pretraining data beyond web scale yields no consistent improvement (Sec. 4.7, Fig. 3). This trend is also evident among VLM vision towers, where transfer performance tracks the degree of proximity to the original contrastive initialization, from the lightly tuned CLIP tower in LLaVA-NeXT (19.8%) to the heavily adapted Qwen towers (32.4–38.6%).

5 Conclusion

We presented a unified benchmark for face PAD, combining LoRA adaptation across 32 vision encoders with zero-shot evaluation of nine complete VLMs under a common protocol on four datasets and against specialist baselines DeepPixBiS and FSFM-FAS.

Our results highlight several key observations. LoRA adaptation substantially improves intra-dataset performance. Most adapted backbones achieve intra-dataset ACER below 2% (Tab. 3), whereas zero-shot prompting of the same models remains near chance (Tab. 2). Cross-dataset transfer remains challenging after adaptation: ACER ranges from 20–43% and is strongly influenced by the choice of training dataset (Tab. 4). Moreover, increasing model scale alone does not consistently improve transferability, as seen with InternViT-6B. The vision encoder of DeepSeek-VL2 contains SigLIP that is originally trained on ImageNet-scale, but in combination in multi-stage training in a whole multimodal model, its performance has increased. Most importantly, the rather small CLIP ViT-B/32 and the much larger extracted CLIP-vision tower from the MLLM LLaVA-NeXT-7B achieve the strongest detection accuracy, whereas CLIP ViT-B/32 outperforms in overall compute Fig. 2.

These findings raise concerns regarding the effectiveness of previous approaches [29, 64] that utilize additional data, such as leave-one-out (LOO) or auxiliary datasets, to enhance generalization. Focusing on other post-training methods, including additional fine-tuning with domain-invariant losses or alignment with target-domain data, may mitigate cross-dataset issues.

References

1. Anjos, A., El Shafey, L., Wallace, R., Günther, M., McCool, C., Marcel, S.: Bob: a free signal processing and machine learning toolbox for researchers. In: 20th ACM Conference on Multimedia Systems (ACMMM) (Oct 2012), https://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf
2. Anjos, A., Günther, M., de Freitas Pereira, T., Korshunov, P., Mohammadi, A., Marcel, S.: Continuously reproducing toolchains in pattern recognition and machine learning experiments. In: International Conference on Machine Learning (ICML) (Aug 2017), http://publications.idiap.ch/downloads/papers/2017/Anjos_ICML2017-2_2017.pdf
3. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning. In: European conference on computer vision. pp. 456–473. Springer (2022)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. CoRR **abs/2502.13923** (2025). https://doi.org/10.48550/ARXIV.2502.13923, https://doi.org/10.48550/arXiv.2502.13923
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
6. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: 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). pp. 612–618. IEEE (2017)
7. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1578–1587 (2022)
8. Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: COYO-700M: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset (2022)
9. Cai, R., Soh, C.Y., Yu, Z., Li, H., Yang, W., Kot, A.C.: Towards data-centric face anti-spoofing: Improving cross-domain generalization via physics-based data synthesis. International Journal of Computer Vision pp. 1–22 (2024)
10. Cao, J., Ma, C.: Towards generalized face anti-spoofing from a frequency shortcut view. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1005–1015. IEEE (2025)
11. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
12. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury,

J., Kuo, W.: Pali: A jointly-scaled multilingual language-image model. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/forum?id=mWVoBz4W0u

13. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
14. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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. pp. 24185–24198 (2024)
15. Chen, Z., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Huang, F., Jin, X.: Generalizable representation learning for mixture domain face anti-spoofing. In: AAAI. pp. 1132–1139. AAAI Press (2021)
16. Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., Marcel, S.: The replay-mobile face presentation-attack database. In: 2016 international conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–7. IEEE (2016)
17. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformer needs registers. In: International Conference on Learning Representations (2024)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
19. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
20. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A.T., Shankar, V.: Data filtering networks. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=KAk6ngZ09F
21. Fang, M., Damer, N.: 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. pp. 6269–6279 (2024)
22. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: 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. pp. 1131–1140. IEEE (2022). https://doi.org/10.1109/WACV51458.2022.00120, https://doi.org/10.1109/WACV51458.2022.00120
23. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Real masks and spoof faces: On the masked face presentation attack detection. Pattern Recognit. **123**, 108398 (2022)
24. Fang, M., Huber, M., Damer, N.: 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. pp. 1061–1070 (2023)
25. Feng, M., Gallin-Martel, P., Ito, K., Aoki, T.: Benchmarking vision foundation models for domain-generalizable face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1567–1576 (2026)

26. de Freitas Pereira, T., Anjos, A., Martino, J.M.D., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: ICB. pp. 1–8. IEEE (2013)
27. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S.M., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A.J., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: Datacomp: In search of the next generation of multimodal datasets. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/56332d41d55ad7ad8024aac625881be7-Abstract-Datasets_and_Benchmarks.html
28. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. In: 2019 international conference on biometrics (ICB). pp. 1–8. IEEE (2019)
29. Gonzalez-Soler, L.J., Tapia, J.E., Busch, C.: 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). pp. 1–10. IEEE (2025)
30. Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., Yao, L., Huang, R., Zhang, W., Jiang, X., Xu, C., Xu, H.: Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/a90b9a09a6ee43d6631cf42e225d73b4-Abstract-Datasets_and_Benchmarks.html
31. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
32. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
33. Huang, H.P., Sun, D., Liu, Y., Chu, W.S., Xiao, T., Yuan, J., Adam, H., Yang, M.H.: Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In: European conference on computer vision. pp. 37–54. Springer (2022)
34. International Organization for Standardization: ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting (2017)
35. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: CVPR. pp. 8481–8490. Computer Vision Foundation / IEEE (2020)
36. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3992–4003. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00371, https://doi.org/10.1109/ICCV51070.2023.00371

37. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
38. Komaty, A., Otroshi-Shahreza, H., George, A., Marcel, S.: Exploring chatgpt for face presentation attack detection in zero and few-shot in-context learning. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025 - Workshops, Tucson, AZ, USA, February 28 - March 4, 2025. pp. 1602–1611. IEEE (2025). https://doi.org/10.1109/WACVW65960.2025.00093, https://doi.org/10.1109/WACVW65960.2025.00093
39. Koupaee, M., Wang, W.Y.: Wikihow: A large scale text summarization dataset. CoRR **abs/1810.09305** (2018), http://arxiv.org/abs/1810.09305
40. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/e2cfb719f58585f779d0a4f9f07bd618-Abstract-Datasets_and_Benchmarks.html
41. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
42. Li, H., Li, W., Cao, H., Wang, S., Huang, F., Kot, A.C.: Unsupervised domain adaptation for face anti-spoofing. IEEE Trans. Inf. Forensics Secur. **13**(7), 1794–1809 (2018)
43. Li, Z., Zhao, T., Xu, X., Zhang, Z., Li, Z., Chen, X., Zhang, Q., Bergamo, A., Jain, A.K., Xing, Y.: Optimal transport-guided source-free adaptation for face anti-spoofing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24351–24363 (2025)
44. Lin, K.H., Tseng, Y.W., Huang, K.Y., Wu, J.C., Cheng, W.H.: Instructflip: Exploring unified vision-language model for face anti-spoofing. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2987–2996 (2025)
45. Liu, A., Xue, S., Gan, J., Wan, J., Liang, Y., Deng, J., Escalera, S., Lei, Z.: Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 222–232 (2024)
46. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/ (2024)
47. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: CVPR. pp. 389–398. IEEE Computer Society (2018)
48. Liu, Y., Chen, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Causal intervention for generalizable face anti-spoofing. In: ICME. pp. 1–6. IEEE (2022)
49. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
50. Otroshi Shahreza, H., Marcel, S.: Foundation models and biometrics: A survey and outlook. IEEE Transactions on Information Forensics and Security (2025).

https://doi.org/10.1109/TIFS.2025.3602233, https://ieeexplore.ieee.org/document/11137396
51. Ozgur, G., Caldeira, E., Chettaoui, T., Boutros, F., Ramachandra, R., Damer, N.: Foundpad: Foundation models reloaded for face presentation attack detection. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 745–755 (2025)
52. Pasmino, D., Aravena, C., Tapia, J.E., Busch, C.: Flickr-pad: New face high-resolution presentation attack detection database. In: 2023 11th International Workshop on Biometrics and Forensics (IWBF). pp. 1–6. IEEE (2023)
53. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE transactions on information forensics and security **11**(10), 2268–2283 (2016)
54. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
55. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: an open large-scale dataset for training next generation image-text models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html
56. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: CVPR. pp. 10023–10031. Computer Vision Foundation / IEEE (2019)
57. Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: AAAI. pp. 11974–11981. AAAI Press (2020)
58. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
59. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021. pp. 2443–2449. ACM (2021). https://doi.org/10.1145/3404835.3463257, https://doi.org/10.1145/3404835.3463257
60. Srivatsan, K., Naseer, M., Nandakumar, K.: Flip: Cross-domain face anti-spoofing with language guidance. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19685–19696 (2023)
61. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
62. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: EVA-CLIP-18B: scaling CLIP to 18 billion parameters. CoRR **abs/2402.04252** (2024). https://doi.org/10.48550/ARXIV.2402.04252, https://doi.org/10.48550/arXiv.2402.04252
63. Wang, C., Lu, Y., Yang, S., Lai, S.: Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In: CVPR. pp. 20249–20258. IEEE (2022)

64. Wang, G., Lin, F., Wu, T., Liu, Z., Ba, Z., Ren, K.: Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24364–24376 (2025)
65. Wang, G., Han, H., Shan, S., Chen, X.: Improving cross-database face presentation attack detection via adversarial domain adaptation. In: ICB. pp. 1–8. IEEE (2019)
66. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. CoRR **abs/2409.12191** (2024). https://doi.org/10.48550/ARXIV.2409.12191, https://doi.org/10.48550/arXiv.2409.12191
67. Wang, Z., Wang, Q., Deng, W., Guo, G.: Face anti-spoofing using transformers with relation-aware mechanism. IEEE Trans. Biom. Behav. Identity Sci. **4**(3), 439–450 (2022)
68. Watanabe, K., Ito, K., Aoki, T.: 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. pp. 1561–1565 (2022)
69. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015)
70. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023)
71. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
72. Yan, W., Zeng, Y., Hu, H.: Domain adversarial disentanglement network with cross-domain synthesis for generalized face anti-spoofing. IEEE Trans. Circuits Syst. Video Technol. **32**(10), 7033–7046 (2022)
73. Yang, J., Lin, X., Yu, Z., Zhang, L., Liu, X., Li, H., Yuan, X., Cao, X.: Dadm: Dual alignment of domain and modality for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12045–12056 (2025)
74. Yu, Z., Li, X., Shi, J., Xia, Z., Zhao, G.: Revisiting pixel-wise supervision for face anti-spoofing. IEEE Trans. Biom. Behav. Identity Sci. **3**(3), 285–295 (2021)
75. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: NAS-FAS: static-dynamic central difference network search for face anti-spoofing. IEEE Trans. Pattern Anal. Mach. Intell. **43**(9), 3005–3023 (2021)
76. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: CVPR. pp. 5294–5304. IEEE (2020)
77. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
78. Zhang, G., Wang, K., Yue, H., Liu, A., Zhang, G., Yao, K., Ding, E., Wang, J.: Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)

79. Zhang, H., Fang, Z., Zhao, N., Hou, S., Ma, L., Pei, R., He, Z.: Harnessing chain-of-thought reasoning in multimodal large language models for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 33566–33576 (June 2026)
80. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
81. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European conference on computer vision. pp. 70–85. Springer (2020)