PlainProbe: A Stable Cross-Entropy Baseline for Data-Scarce Deepfake Detection

Anonymous Authors[1]

Abstract

Deepfake detectors are difficult to stabilize under new generators and source distributions. We present an objective-ablation framework for adapting deepfake detectors with limited data. Using a fixed GenD-based CLIP detector, we reformulate training as standard cross-entropy over real and fake labels by removing auxiliary alignment and uniformity objectives. Our empirical results show that this simplified objective yields more stable training, reducing AUROC standard deviation across five random seeds by 57–75% relative to a matched GenD-based reference. On Celeb-DF v2, FFTEST, and WILDD, our method achieves AUROC/mAP scores of 0.7287/0.7000, 0.9867/0.9857, and 0.8410/0.8457, respectively. These results suggest that stable, reproducible objectives can serve as useful baselines for reliability-oriented deepfake detection research.

1. Introduction

Deepfake detection is increasingly evaluated beyond in-distribution binary classification on fixed benchmarks alone. In real deployment, manipulation cues shift with generators, editing pipelines, compression settings, platform conditions, and preprocessing protocols. As a result, strong performance on one benchmark does not guarantee robust performance on newly emerging fake-media sources.

This gap matters most in early-response scenarios, where detectors must adapt before larger validation sets become available.

Prior work has advanced deepfake detection through public benchmarks (Rössler et al., 2019; Li et al., 2020; Zi et al., 2020; Zhou et al., 2021; Yan et al., 2023) and parameter-efficient adaptation of CLIP-based encoders (Radford et al., 2021; Hu et al., 2022; Ben Zaken et al., 2022; Yermakov et al., 2026), often augmented with auxiliary objectives such as alignment and uniformity (Wang & Isola, 2020).

However, benchmark generalization does not directly translate to operational adaptability. When a new fake-media source appears, building a large, balanced validation set is rarely immediate. Early response is typically constrained to a low-data setting with a small number of verified samples. Under this constraint, the key question is not merely mean performance, but also seed stability, defined here as low seed-to-seed variability under fixed data lists and video-level grouping rules.

Motivated by this, we investigate whether auxiliary alignment and uniformity objectives improve stability under a fixed low-data protocol. To do so, we isolate the training objective in the GenD-based reference (GenD-Ref) of Yermakov et al. (2026), with all other components fixed.

We refer to this isolated variant as PlainProbe**,** the plain cross-entropy version of GenD-Ref that shares the same architecture (CLIP ViT-L/14 with LayerNorm adaptation, L2 normalization, and a linear classifier) but removes the alignment and uniformity terms. The comparison is thus a controlled objective ablation, not a new-architecture comparison.

Across Celeb-DF v2 (CDFv2), an FF++-aligned held-out test set (FFTEST), and a WildDeepfake-derived evaluation set (WILDD), PlainProbe provides the best aggregate mean-stability trade-off under the fixed protocol. Frozen linear probing alone is insufficient, while LoRA (low-rank adaptation) and BitFit (bias fine-tuning) are competitive on some benchmarks but less consistent across the full suite. Some adaptation is therefore necessary, but auxiliary alignment/uniformity terms do not consistently improve stability in this low-data regime. However, this aggregate advantage does not extend to strict low-FPR operating points, where GenD-Ref can remain preferable on certain distributions, motivating regime-specific recommendations rather than a single aggregate winner.

All claims pertain to the defined protocol and GenD-Ref setting, within which standard cross-entropy training serves as a stable, reproducible baseline.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Our contributions are as follows:

- An objective-ablation baseline, PlainProbe, that removes auxiliary alignment/uniformity losses from the matched GenD-Ref and trains the same architecture with standard cross-entropy.
- Empirical evidence over five random seeds on Celeb-DF v2 (CDFv2), an FF++-aligned held-out test set (FFTEST), and a WildDeepfake-derived evaluation set (WILDD), showing that PlainProbe reduces seed-to-seed AUROC standard deviation by 57-75% relative to the GenD-Ref while maintaining competitive mean AUROC/mAP.
- A reliability-oriented analysis arguing that detector evaluation should jointly consider mean performance, seed stability, low-FPR operating points, source-list construction, grouping rules, and trainable-parameter footprint, rather than average metrics alone.

2. Related Work

2.1. Pretrained Visual Encoders and Lightweight Adaptation

Recent deepfake detection research has increasingly adopted large-scale pretrained visual encoders. In particular, vision-language models such as CLIP provide transferable representations learned from large corpora, and have become strong backbones for face forgery and deepfake detection (Radford et al., 2021; Khan & Dang-Nguyen, 2024; Cui et al., 2025; Wang et al., 2025).

Because full fine-tuning of large encoders is expensive, lightweight adaptation methods that update a small subset of parameters are now widely used. Methods such as LoRA (low-rank adaptation), BitFit (bias fine-tuning), and LayerNorm adaptation reduce adaptation cost by freezing most backbone parameters while training limited components (Hu et al., 2022; Ben Zaken et al., 2022). GenD is closely aligned with this trend. It combines CLIP ViT-L/14, LayerNorm adaptation, L2-normalized features, and a linear classifier, while augmenting standard cross-entropy with alignment and uniformity losses to improve cross-dataset generalization (Yermakov et al., 2026).

2.2. Cross-Dataset Generalization and Benchmark Sensitivity

Cross-dataset evaluation has become a core protocol in deepfake detection. Public benchmarks such as FaceForensics++, Celeb-DF, WildDeepfake, FFIW, and DeepfakeBench cover diverse manipulation types, identities, capture conditions, compression settings, and preprocessing pipelines, providing standardized grounds for assessing generalization (Rössler et al., 2019; Li et al., 2020; Zi et al., 2020; Zhou et al., 2021; Yan et al., 2023).

However, cross-dataset performance can be sensitive to evaluation details. Each benchmark has its own source videos, identity distribution, compression setting, face detection pipeline, frame sampling strategy, score aggregation rule, and video-level grouping policy. As a result, reported results may reflect not only model capability but also data-list construction and evaluation-code choices. This protocol sensitivity makes detector reliability difficult to summarize with a single average metric.

2.3. Auxiliary Representation-Learning Objectives

A variety of studies have explored auxiliary representation-learning objectives to supplement standard classification loss and improve generalization. Alignment and uniformity objectives encourage better separation in normalized feature space by regulating sample relations (Wang & Isola, 2020). Prototype-based and subspace-based approaches also attempt to structure feature geometry to reduce overfitting to specific manipulation sources (Yan et al., 2024; 2025).

These objectives can be effective when training data are diverse and pair construction is reliable. In low-data settings, however, positive-pair definitions, source grouping, class balance, and mini-batch composition may dominate the learning signal. In such cases, auxiliary objectives may produce representations that are more sensitive to local data-list and batch composition than to transferable manipulation cues. This makes it important to isolate auxiliary objectives under matched model configurations, especially for low-data adaptation.

3. Method

3.1. Overview

Our aim is not to propose a new detector architecture, but to isolate the effect of the training objective within a GenD-Ref setting. We adopt the same GenD-Ref components (CLIP ViT-L/14, LayerNorm adaptation, L2-normalized features, and a linear classifier) and isolate the objective by removing alignment/uniformity terms, as illustrated in Figure 1. We focus on frame-level CLIP-based detection. Video-centric and multimodal systems are beyond our scope.

3.2. Model Architecture

Let an input face crop be denoted by x. PlainProbe first extracts an image feature using the CLIP ViT-L/14 encoder. Most CLIP parameters are frozen, while the affine parameters of LayerNorm are trainable:

$$z = f_{\theta_0, \theta_{\mathrm{LN}}}(x) \in \mathbb{R}^d, \tag{1}$$

where θ_0 denotes frozen CLIP parameters and θ_{LN} denotes trainable LayerNorm affine parameters. This setup enables

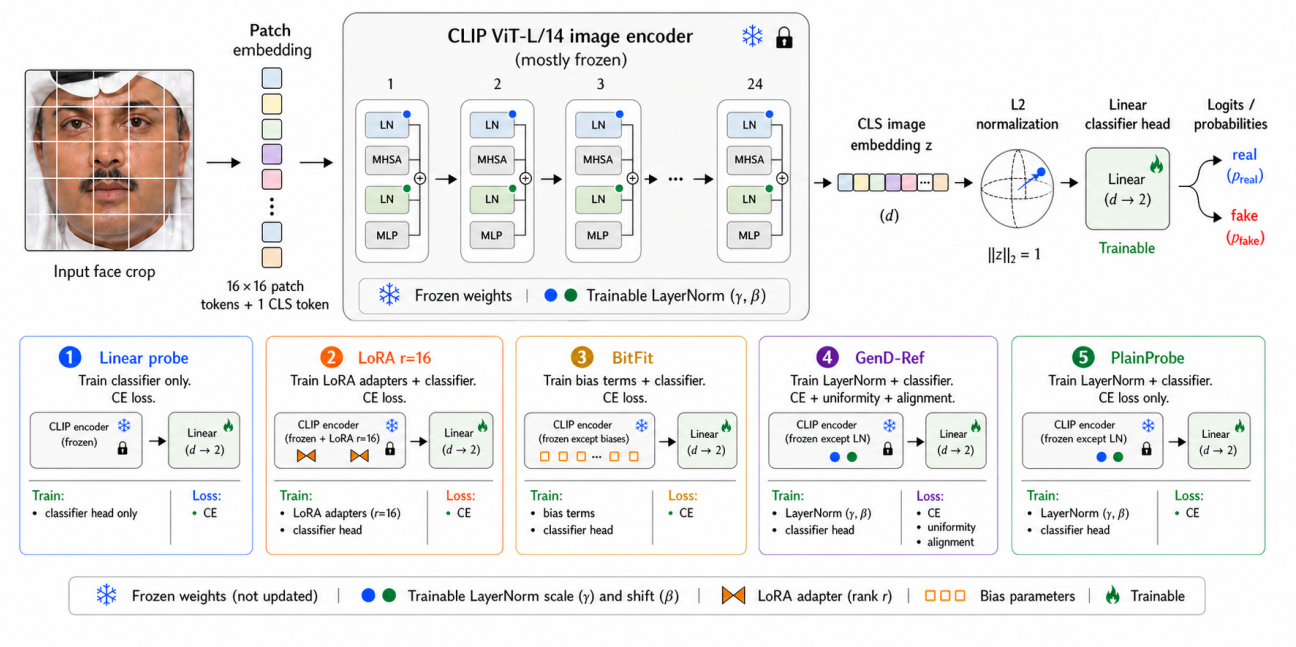

Figure 1. GenD-based detector variants with different trainable parts and objectives (all use CLIP ViT-L/14).

limited adaptation of pretrained representations without full-backbone fine-tuning.

The extracted feature is then L2-normalized:

$$\tilde{z} = \frac{z}{\|z\|_2 + \epsilon}, \tag{2}$$

where ϵ is a small constant for numerical stability. This normalization encourages the classifier to focus on feature direction rather than raw magnitude.

The normalized feature $\tilde{z}$ is passed to a binary linear classifier:

$$\ell = W\tilde{z} + b, \tag{3}$$

where $\ell \in \mathbb{R}^2$ contains the logits for real and fake classes. We use $y = 0$ for real and $y = 1$ for fake. The frame-level fake probability is given by

$$p_\phi(y = 1 \mid x) = [\mathrm{softmax}(\ell)]_{\mathrm{fake}}, \tag{4}$$

where $\phi = (W, b)$ are classifier parameters. The trainable parameters are therefore θ_{LN} and $\phi = (W, b)$.

3.3. Training Objective

PlainProbe removes the auxiliary alignment and uniformity losses used in the GenD-Ref and is trained with standard binary cross-entropy over observed real/fake labels. This direct classification objective avoids positive-pair construction, source-pair construction, and mini-batch feature-distribution estimation. We hypothesize that, in low-data regimes, removing these auxiliary dependencies reduces sensitivity to pair definitions and batch composition while preserving the direct real/fake classification signal.

3.4. Controlled Comparison

The controlled comparison fixes the CLIP ViT-L/14 backbone, LayerNorm adaptation, L2-normalized features, and the linear classifier. The intended difference between PlainProbe and the GenD-Ref is the training objective, since the GenD-Ref includes alignment and uniformity terms whereas PlainProbe uses standard cross-entropy only. Thus, all reported differences between PlainProbe and the GenD-Ref should be interpreted as objective-level effects under a fixed architecture and protocol, rather than architectural gains.

We use the GenD-Ref with its original default coefficients for the alignment and uniformity terms, and do not tune them per dataset. This is intentional. Under our low-data adaptation premise, dataset-specific coefficient tuning would itself constitute supervision over the held-out distribution, contradicting our early-response setting. The comparison therefore reflects how each objective behaves under fixed default configurations, which is the realistic deployment regime when verified samples are scarce.

To control for adaptation capacity, we also compare frozen linear probing, BitFit, and LoRA ($r = 16$).

4. Experimental Setup

In this section, we describe the experimental setup used to evaluate the stability of PlainProbe. We evaluate CDFv2 and

WILDD as primary public OOD benchmarks, and FFTEST as an FF++-aligned in-protocol stress check. Our evaluation has three goals. First, under a fixed GenD-based detector setting, we analyze how objective choice affects cross-dataset transfer and seed stability. Second, we evaluate not merely mean AUROC and mAP, but also behavior under low false-positive-rate regimes. Third, we fix data lists, video-level grouping rules, evaluation code, and random seeds so that observed differences are attributable to adaptation strategy and objective choice rather than protocol drift.

Our main empirical claim is limited to the defined low-data protocol and the main evaluation suite (CDFv2, WILDD, FFTEST). Within this scope, we evaluate whether PlainProbe provides a stable balance between mean transfer performance and seed stability.

4.1. Datasets

We evaluate the main variants on three video-level test sets: CDFv2, WILDD, and FFTEST. CDFv2 follows Celeb-DF v2, treating Celeb-synthesis as fake and Celeb-real/YouTube-real as real (Li et al., 2020). WILDD is derived from WildDeepfake and contains more unconstrained source variation (Zi et al., 2020). FFTEST is an FF++-based held-out list aligned with the training source (Rössler et al., 2019).

4.2. Video-Level Grouping

For all experiments, frames are grouped into videos using a canonical video identifier parsed from the frame filename. This approach avoids aggregation artifacts caused by inconsistent directory structures, where grouping by parent directory may either merge distinct videos or split frames belonging to the same video. For CDFv2, video identifiers are reconstructed from filenames rather than inherited from parent directories.

4.3. Compared Variants

We compare PlainProbe with controlled variants under the same data lists, training protocol, and video-level evaluation pipeline. The variants are selected to separate the effect of adaptation capacity from the effect of the training objective.

Frozen linear probing. Only the final classifier is trained, while the CLIP ViT-L/14 encoder remains frozen. This lower-bound sanity check evaluates whether frozen CLIP features alone are sufficient under the low-data protocol.

BitFit. Following BitFit (Ben Zaken et al., 2022), we train the bias parameters together with the classifier, keeping the remaining backbone parameters frozen. This provides a lightweight adaptation baseline with a larger trainable subset than linear probing and a smaller trainable footprint than LoRA.

LoRA ($r = 16$). Rank-16 low-rank adapters are added to the query, key, value, and output projections in each self-attention layer (Hu et al., 2022). The pretrained weights remain frozen, and only the low-rank adapter parameters and classifier are trained. This variant tests whether increasing parameter-efficient adaptation capacity improves low-data transfer.

GenD-based reference. The GenD-Ref uses the same CLIP ViT-L/14 backbone, LayerNorm adaptation, L2-normalized features, and linear classifier as PlainProbe. It is trained with cross-entropy plus auxiliary alignment and uniformity objectives (Yermakov et al., 2026).

PlainProbe. PlainProbe uses the same backbone, trainable components, feature normalization, and classifier as the GenD-Ref, but removes the auxiliary alignment and uniformity objectives. It is trained only with standard cross-entropy over real/fake labels. Therefore, this comparison isolates the effect of the training objective under matched model components.

4.4. Training Protocol

All main comparisons are conducted under the same FF++ low-data adaptation protocol. The training split contains 480 fake and 480 real frames, and the validation split contains 60 fake and 60 real frames. This 960-frame training budget reflects an early adaptation setting with limited verified labels.

We use CLIP ViT-L/14 as the backbone and resize all inputs to the CLIP resolution of 224×224. For the main evaluation suite, face crops are extracted using a DeepfakeBench-compatible preprocessing pipeline (Yan et al., 2023). All variants use the same training and validation lists.

All models are optimized with AdamW. The learning rate follows a cosine decay schedule from 3×10^{-4} to 1×10^{-5}, with one warmup epoch. Training is performed for 50 epochs. We use gradient accumulation to obtain an effective batch size of 32. All runs use mixed-precision training and four data-loading workers.

4.5. Frame-Level and Video-Level Evaluation

For each sampled frame x_i, the detector outputs a frame-level fake probability $p_\phi(y = 1 \mid x_i)$. Frame scores are averaged over the sampled frames F_v of each video to obtain a video-level fake score:

$$s_v = \frac{1}{|F_v|} \sum_{x_i \in F_v} p_\phi(y = 1 \mid x_i). \tag{5}$$

Table 1. Video-level AUROC and mAP (mean ± SD, 5 seeds) on CDFv2, FFTEST, and WILDD. Frozen linear probing is included as a lower-bound sanity check. Best means in bold; †: selected aggregate baseline.

	CDFv2		FFTEST		WILDD	
Method	AUROC	mAP	AUROC	mAP	AUROC	mAP
Frozen linear probe	0.4526 ± 0.0024	0.4812 ± 0.0018	0.7800 ± 0.0031	0.7942 ± 0.0030	0.6428 ± 0.0051	0.6398 ± 0.0052
LoRA r16	0.7437 ± 0.0656	0.7217 ± 0.0675	0.9425 ± 0.0257	0.9407 ± 0.0257	0.7586 ± 0.0379	0.7694 ± 0.0342
BitFit	**0.7559 ± 0.0294**	**0.7371 ± 0.0223**	0.9658 ± 0.0152	0.9641 ± 0.0186	0.7725 ± 0.0579	0.7674 ± 0.0650
GenD-Ref	0.7105 ± 0.0558	0.6821 ± 0.0476	0.9517 ± 0.0308	0.9496 ± 0.0318	0.8145 ± 0.0213	0.8222 ± 0.0189
PlainProbe†	0.7287 ± 0.0141	0.7000 ± 0.0156	**0.9867 ± 0.0122**	**0.9857 ± 0.0133**	**0.8410 ± 0.0091**	**0.8457 ± 0.0085**

All reported metrics are computed from these video-level scores using the same aggregation rule across methods and datasets.

4.6. Metrics

We report video-level AUROC and mean average precision (mAP) as threshold-independent metrics. For each method, results are reported as mean ± standard deviation over five random seeds. Seed-to-seed standard deviation is used as a descriptive measure of stability under the low-data adaptation protocol. Given the small sample size (n = 5), we do not perform formal hypothesis testing on variance differences; we instead rely on consistent ordering across all three benchmarks as the primary stability evidence.

For operating-point analysis, we additionally report TPR at fixed FPRs of 0.1%, 1%, and 5%. Because these values are computed on finite video-level benchmarks, the attainable FPR grid depends on the number of real videos. This effect is most visible on FFTEST, where the limited number of negative videos yields a coarse FPR grid. Consequently, the reported TPR values coincide across the three FPR budgets.

5. Results and Discussion

We analyze how the training objective affects low-data adaptation under the fixed protocol described above. CDFv2 and WILDD are used as the main public cross-dataset evaluations, while FFTEST serves as an FF++-aligned held-out evaluation. Unless otherwise stated, results are reported at the video level over five random seeds.

5.1. Main Evaluation Suite

Table 1 reports video-level AUROC and mAP as mean ± standard deviation over five random seeds. We assess methods by considering both mean performance and seed-to-seed variation, since average AUROC/mAP alone can obscure instability in low-data adaptation. The primary controlled comparison is between PlainProbe and the GenD-Ref, which use matched model components but different training objectives.

Under this fixed protocol, PlainProbe achieves higher mean AUROC/mAP and lower seed-to-seed variation than the GenD-Ref on CDFv2, FFTEST, and WILDD. For AUROC, the standard deviation decreases from 0.0558 to 0.0141 on CDFv2, from 0.0308 to 0.0122 on FFTEST, and from 0.0213 to 0.0091 on WILDD. This reduction is important in the low-data setting, where random initialization and mini-batch ordering can have amplified effects. The corresponding seed-level distributions are visualized in Figure 2, highlighting the reduced seed-to-seed variation of PlainProbe relative to the matched GenD-Ref.

The control variants clarify the role of adaptation capacity. Frozen linear probing yields substantially lower AUROC on all three benchmarks, suggesting that some adaptation of the pretrained representation is necessary. LoRA and BitFit are competitive on CDFv2, with BitFit achieving the highest CDFv2 mean AUROC/mAP. However, these variants are less consistent across the full suite: LoRA exhibits larger seed-to-seed variation on CDFv2, and BitFit exhibits larger seed-to-seed variation on WILDD.

Overall, PlainProbe provides the most consistent aggregate behavior across the main evaluation suite. Although it does not achieve the best mean result on every individual benchmark, it offers a favorable balance between mean performance and seed stability under the fixed low-data protocol.

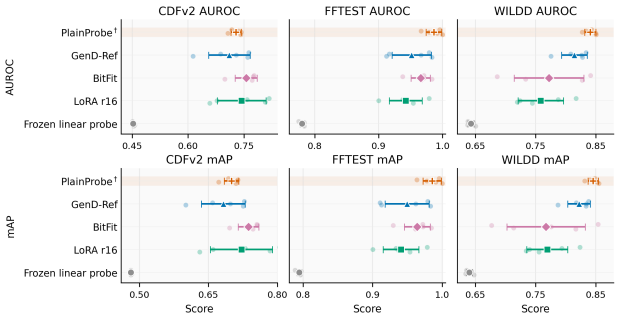

Figure 2. Visualization of video-level AUROC and mAP across five random seeds on CDFv2, FFTEST, and WILDD. Each dot denotes one seed result, the central marker denotes the mean, and horizontal bars denote standard deviation. PlainProbe exhibits consistently lower seed-to-seed variation than the matched GenD-Ref across the main evaluation suite.

Table 2. Ablation of LayerNorm adaptation (LN), L2 normalization, and auxiliary objectives (Aux.) on AUROC/mAP (mean ± SD, 5 seeds).

LN	L2	Aux.	CDFv2		FFTEST		WILDD	
			AUROC	mAP	AUROC	mAP	AUROC	mAP
✓	✗	✗	0.6788 ± 0.0228	0.6535 ± 0.0202	0.9467 ± 0.0076	0.9447 ± 0.0096	0.7795 ± 0.0195	0.7742 ± 0.0276
✓	✓	✗	$\mathbf{0.7287 \pm 0.0141}$	$\mathbf{0.7000 \pm 0.0156}$	$\mathbf{0.9867 \pm 0.0122}$	$\mathbf{0.9857 \pm 0.0133}$	$\mathbf{0.8410 \pm 0.0091}$	$\mathbf{0.8457 \pm 0.0085}$
✓	✓	✓	0.7105 ± 0.0558	0.6821 ± 0.0476	0.9517 ± 0.0308	0.9496 ± 0.0318	0.8145 ± 0.0213	0.8222 ± 0.0189

5.2. Ablation on Normalization and Auxiliary Objectives

Table 2 isolates the effects of L2 feature normalization and the auxiliary alignment/uniformity objectives. Adding L2 normalization to LayerNorm adaptation improves AUROC and mAP on all three main benchmarks. For AUROC, performance increases from 0.6788 to 0.7287 on CDFv2, from 0.9467 to 0.9867 on FFTEST, and from 0.7795 to 0.8410 on WILDD. These gains suggest that feature normalization is an important component of the GenD-based CLIP setting.

By contrast, adding alignment/uniformity objectives on top of the L2-normalized LayerNorm setting does not improve mean AUROC or mAP in this diagnostic study. It also increases seed-to-seed variation relative to the L2-only setting across the three benchmarks.

5.3. Low-FPR Operating Behavior

Table 3 reports TPR at 0.1%, 1%, and 5% FPR. These results show that conclusions should not be based solely on mean AUROC/mAP. On CDFv2, the GenD-Ref has higher TPR than PlainProbe at all three FPR budgets. On WILDD, the GenD-Ref is stronger at 0.1% and 1% FPR, while PlainProbe is stronger at 5% FPR. On FFTEST, the limited number of negative videos collapses the achievable FPR grid such that all three reported budgets correspond to a single operating point. We therefore exclude FFTEST from the low-FPR comparison and report its values for completeness only.

Accordingly, our claims are bounded by this operating-point behavior. Although PlainProbe is the selected aggregate baseline in Table 1, the preferred method can change under strict false-positive-rate budgets. On both CDFv2 and WILDD, GenD-Ref achieves higher TPR than PlainProbe at most strict-FPR operating points. In deployment scenarios where low-FPR constraints are paramount, GenD-Ref therefore remains the preferred objective. PlainProbe's advantage is restricted to aggregate threshold-independent metrics and seed stability.

5.4. Trainable-Parameter Footprint

Table 4 summarizes the trainable-parameter footprint of each adaptation strategy. LayerNorm adaptation settings, including PlainProbe and the GenD-Ref, train 104,450 parameters, corresponding to 0.0344% of the detector. These two methods have identical trainable-parameter counts, as they share the same trainable components and differ only in the training objective.

By contrast, LoRA trains 3,147,778 parameters, corresponding to 1.0250% of the detector. Thus, PlainProbe uses roughly $30\times$ fewer trainable parameters than LoRA.

Frozen linear probing yields the smallest trainable footprint but offers limited adaptation capacity in this setting. LoRA provides substantially larger adaptation capacity but does not consistently improve seed stability across the main evaluation suite. Overall, PlainProbe retains a small trainable footprint while providing stable aggregate performance under the fixed low-data protocol. This makes it a useful baseline when both trainable-parameter cost and verified-data availability are constrained.

5.5. Discussion

Under the fixed low-data protocol, PlainProbe provides the most stable internal baseline when mean performance and seed-to-seed variation are considered jointly. Our results suggest that, under this specific protocol and pair construction, the benefits of alignment/uniformity objectives may be outweighed by increased sensitivity to mini-batch composition and random seeds.

A plausible explanation is sensitivity in the low-data regime. With limited verified labels, pair sampling and mini-batch composition may dominate the auxiliary signal, whereas cross-entropy directly updates the real/fake decision boundary from observed labels. This interpretation is consistent with the ablation in Table 2, where adding alignment/uniformity terms on top of the L2-normalized LayerNorm setting does not improve mean AUROC or mAP and increases seed-to-seed variation.

This pattern is consistent with the possibility that pair-based losses amplify the influence of mini-batch composition when training data are limited. Under our 480/480 training budget, each mini-batch covers only a small fraction of source variation. Cross-entropy is computed per sample and avoids this batch-level dependency, which may help explain the observed stability gap.

Table 3. TPR at 0.1%, 1%, and 5% FPR (video-level). For FFTEST, identical values result from few negative videos.

Benchmark	Method	TPR@0.1% FPR	TPR@1% FPR	TPR@5% FPR
CDFv2	LoRA r16	0.0382 ± 0.0369	0.1247 ± 0.0951	0.2665 ± 0.0827
	BitFit	0.0935 ± 0.0633	0.1212 ± 0.0645	0.2753 ± 0.0770
	GenD-Ref	**0.2571 ± 0.0595**	**0.2641 ± 0.0627**	**0.3794 ± 0.0943**
	PlainProbe	0.1482 ± 0.0384	0.1835 ± 0.0192	0.3306 ± 0.0297
FFTEST	LoRA r16	0.8300 ± 0.0245	0.8300 ± 0.0245	0.8300 ± 0.0245
	BitFit	0.8200 ± 0.0510	0.8200 ± 0.0510	0.8200 ± 0.0510
	GenD-Ref	0.8300 ± 0.0245	0.8300 ± 0.0245	0.8300 ± 0.0245
	PlainProbe	**0.9400 ± 0.0374**	**0.9400 ± 0.0374**	**0.9400 ± 0.0374**
WILDD	LoRA r16	0.0839 ± 0.0173	0.2419 ± 0.0360	0.3760 ± 0.0435
	BitFit	0.1071 ± 0.1075	0.1895 ± 0.1311	0.3416 ± 0.1878
	GenD-Ref	**0.2187 ± 0.0285**	**0.3880 ± 0.0587**	0.5386 ± 0.0352
	PlainProbe	0.1506 ± 0.0434	0.3610 ± 0.0546	**0.6135 ± 0.0272**

Table 4. Trainable parameter count and share for each adaptation method. LayerNorm adaptation covers both PlainProbe and GenD-Ref.

Setting	Trainable params	Share
Linear probe	2,050	0.0007%
BitFit	274,434	0.0903%
LoRA r16	3,147,778	1.0250%
LayerNorm adaptation (incl. PlainProbe)	**104,450**	**0.0344%**

The low-FPR analysis further qualifies this conclusion. On CDFv2 and at some WILDD operating points, the GenD-Ref achieves higher TPR under strict FPR budgets. Thus, objective choice should be matched to the target operating regime. PlainProbe provides stable aggregate behavior, while the GenD-Ref can remain preferable when performance at very low FPRs is the primary criterion.

5.6. Scope and Future Directions

Our findings should be interpreted within the GenD-Ref-based low-data adaptation protocol studied here. We use a matched local GenD-Ref implementation to isolate objective-level effects under fixed model components and evaluation code. We also keep the default auxiliary-loss coefficients without per-benchmark tuning, which matches the early-response setting where verified validation data are scarce. Our evaluation focuses on frame-level CLIP-based detection. Extending the analysis to coefficient sensitivity, temporal and multimodal modeling, and broader stress-test settings remains an important direction.

6. Conclusion

We study deepfake detection under domain shift as a protocol-sensitive reliability problem. Within the specified GenD-based low-data protocol, PlainProbe provides the most favorable balance of mean performance and seed stability across CDFv2, FFTEST, and WILDD among the tested internal variants.

The results also show that aggregate AUROC/mAP alone is insufficient for characterizing detector reliability. Under stricter low-FPR operating points, GenD-Ref achieves higher TPR than PlainProbe on both CDFv2 and WILDD at most strict-FPR operating points and remains the preferred objective when low-FPR constraints are paramount. Thus, objective choice should be evaluated in the context of the target operating regime rather than by average performance alone.

Overall, PlainProbe serves as a simple and reproducible objective-ablation baseline for low-data deepfake detector adaptation. More broadly, our results suggest that reliable evaluation should report not only average performance but also video-level grouping rules, source-list construction, seed-to-seed variability, low-FPR behavior, and trainable-parameter footprint. Future work could investigate whether more principled pair construction or curriculum strategies can make auxiliary objectives more stable under strict low-data budgets. The controlled-comparison framing used here may also inform future evaluation of training objectives in low-data adaptation.

Impact Statement

This work studies reliability-oriented deepfake detection under distribution shift. Deepfake detectors can support media verification, online safety, journalism, and forensic analysis, but their deployment also carries risks related to misuse, overreliance, privacy, and false-positive harm. Accordingly, the reported results should be interpreted as bounded benchmark evidence rather than deployment guarantees. Practical deployment should include score calibration, operating-point selection, human oversight, dataset-bias assessment, privacy safeguards, and transparent communication of uncertainty.

References

Ben Zaken, E., Goldberg, Y., and Ravfogel, S. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.1. URL https://aclanthology.org/2022.acl-short.1/.

Cui, X., Li, Y., Luo, A., Zhou, J., and Dong, J. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217, 2025.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Khan, S. A. and Dang-Nguyen, D.-T. CLIPping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024. doi: 10.1145/3652583.3658035.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

Wang, G., Lin, F., Wu, T., Liu, Z., Ba, Z., and Ren, K. FSFM: A generalizable face security foundation model via self-supervised facial representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24364–24376, 2025.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., and Wu, B. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=hizSx8pf0U.

Yan, Z., Luo, Y., Lyu, S., Liu, Q., and Wu, B. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8984–8994, 2024. doi: 10.1109/CVPR52733.2024.00858.

Yan, Z., Wang, J., Jin, P., Zhang, K.-Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., and Yuan, L. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 2025. URL https://proceedings.mlr.press/v267/yan25b.html.

Yermakov, A., Čech, J., Matas, J., and Fritz, M. Deepfake detection that generalizes across benchmarks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 773–783, 2026.

Zhou, T., Wang, W., Liang, Z., and Shen, J. Face forensics in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5778–5788, 2021.

Zi, B., Chang, M., Chen, J., Ma, X., and Jiang, Y.-G. WildDeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2382–2390, 2020. doi: 10.1145/3394171.3413769.

Table 5. Auxiliary-coefficient sensitivity diagnostic under the GenD-Ref CLIP/LN setting. All rows use the same backbone, trainable components, data lists, and evaluation protocol; only the auxiliary coefficients are varied. Results are video-level AUROC/mAP over three seeds (mean ± SD). Core mean is the unweighted average over CDFv2, FFTEST, and WILDD. This diagnostic is not used for model selection and does not replace the five-seed main comparison in Table 1.

Objective setting	λ_u	λ_a	CDFv2 AUROC/mAP	FFTEST AUROC/mAP	WILDD AUROC/mAP	Core mean AUROC/mAP
PlainProbe (standard cross-entropy only)	0.00	0.00	0.7320 ± 0.0105 / 0.7060 ± 0.0092	0.9806 ± 0.0119 / 0.9791 ± 0.0131	0.8347 ± 0.0064 / 0.8403 ± 0.0069	0.8491 ± 0.0028 / 0.8418 ± 0.0028
Standard cross-entropy + auxiliary terms	0.10	0.02	$\mathbf{0.7587 \pm 0.0072}$ / $\mathbf{0.7284 \pm 0.0066}$	$\mathbf{0.9931 \pm 0.0020}$ / $\mathbf{0.9929 \pm 0.0019}$	$\mathbf{0.8430 \pm 0.0058}$ / $\mathbf{0.8473 \pm 0.0064}$	$\mathbf{0.8649 \pm 0.0029}$ / $\mathbf{0.8562 \pm 0.0031}$
Standard cross-entropy + auxiliary terms	0.25	0.05	0.7302 ± 0.0704 / 0.7042 ± 0.0636	0.9556 ± 0.0423 / 0.9543 ± 0.0430	0.8251 ± 0.0369 / 0.8319 ± 0.0335	0.8369 ± 0.0496 / 0.8301 ± 0.0465
GenD-Ref setting	0.50	0.10	0.7265 ± 0.0308 / 0.6948 ± 0.0281	0.9542 ± 0.0270 / 0.9507 ± 0.0291	0.8230 ± 0.0107 / 0.8300 ± 0.0090	0.8346 ± 0.0227 / 0.8252 ± 0.0219
Standard cross-entropy + auxiliary terms	1.00	0.20	0.6482 ± 0.0332 / 0.6280 ± 0.0252	0.9236 ± 0.0205 / 0.9161 ± 0.0216	0.7962 ± 0.0123 / 0.8044 ± 0.0117	0.7893 ± 0.0217 / 0.7828 ± 0.0193

A. Additional Auxiliary-Coefficient Diagnostic

Auxiliary-coefficient sensitivity. To test whether the behavior of auxiliary alignment/uniformity objectives depends on their weighting, we conduct a post-hoc three-seed coefficient diagnostic in Table 5. The results show that auxiliary objectives are coefficient-sensitive: weak auxiliary weights improve the three-seed core mean, whereas the default GenD-Ref and stronger auxiliary weights reduce performance and increase variability. This diagnostic does not replace the five-seed main comparison and is not used for model selection. Instead, it supports our main interpretation that auxiliary objectives require careful coefficient control under low-data adaptation, while PlainProbe remains a simple and stable baseline without coefficient tuning.