LESS PRIOR, MORE FORENSICS: REDUCING PRETRAINED SEMANTIC PRIORS FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The growing realism of generative models has blurred the boundary between real and synthetic content, posing significant challenges to reliable AI-generated image detection. In this paper, we identify a failure pattern, termed *semantic fallback*, where semantic priors inherited from pretraining remain dominant in the representation space after forensic fine-tuning, thereby limiting the generalization of VFM-based detectors. Building on this insight, we propose a **Geometric Semantic Decoupling (GSD)** framework, which explicitly suppresses dominant pretrained geometry, thereby promoting invariant forensic representations. Specifically, GSD leverages a frozen CLIP encoder to estimate the dominant pretrained subspace via Singular Value Decomposition (SVD). It then suppresses the prior-aligned components through a geometry-constrained formulation with the suppression strength adaptively modulated across samples and layers. We further introduce a mini-batch SVD approximation strategy that amortizes subspace estimation, achieving over a 15× reduction in computational overhead while preserving effectiveness. Empirical results show that our method achieves state-of-the-art performance on various datasets, with average video-level AUCs of 94.5% (+2.6%) and 97.8% (+3.0%) on two face-forgery protocols, and average accuracies of 97.3% (+2.1%) on the synthetic-image benchmark. Our code is available at `https://anonymous.4open.science/r/Null-Space_CLIP`.

1 INTRODUCTION

The rapid progress of generative AI (Midjourney, 2022; Rombach et al., 2022) has made it difficult to distinguish real from artificially generated content, motivating the need for detectors that generalize to rapidly evolving generation techniques (Business, 2024; Brown, 2024). In response, the digital forensic community has undergone a substantial methodological shift, moving from hand-crafted statistical artifact analysis toward deep-learning-based methods. Recent state-of-the-art detectors increasingly adopt large-scale pretrained Vision Foundation Models (VFMs), such as CLIP (Radford et al., 2021), as their visual backbones (Chen et al., 2026; Li et al., 2025; Han et al., 2025; Cai et al., 2025). While these methods have improved detection performance, their generalization to unseen domains remains limited. Existing studies typically attribute this generalization gap to unstable cues acquired during downstream forensic training, such as manipulation-specific artifacts (Li et al., 2025; Shiohara & Yamasaki, 2022) and forgery-irrelevant content biases (Yan et al., 2023a; Fu et al., 2025). However, this perspective primarily focuses on what detectors newly acquire during forensic adaptation, while paying much less attention to an equally important question: *how do the semantic priors inherited from VFM pretraining evolve during downstream forensic adaptation?*

We first analyze feature distributions from a pretrained CLIP encoder and its counterpart fine-tuned on FaceForensics++ (Rossler et al., 2019) using t-SNE (Maaten & Hinton, 2008). As shown in Figure 1, fine-tuned detectors do not fully reorganize their representation space around forensic evidence. Instead, clear identity-related organization as a major manifestation of the inherited semantic priors persists after forensic adaptation. We term this phenomenon *semantic fallback*, where *semantic priors inherited from CLIP pretraining persist as a dominant geometric prior after downstream forensic training*. This raises a critical question: *how does semantic fallback affect the generalization of AI-generated image detectors?* We investigate this question in Section 2. Empirically,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

we find that (a) semantic fallback persists after forensic fine-tuning and becomes more pronounced on unseen domains; (b) detection failures are more strongly associated with greater occupancy in the inherited CLIP semantic subspace; and (c) suppressing components aligned with the pretrained geometry improves cross-domain performance under linear probing. We further provide an analytical view of the target-risk tradeoff, suggesting a non-trivial optimal strength that balances semantic interference against the preservation of discriminative forensic cues. These findings identify semantic fallback as a key factor limiting the generalization of VFM-based detectors, rather than an incidental property of their learned representations.

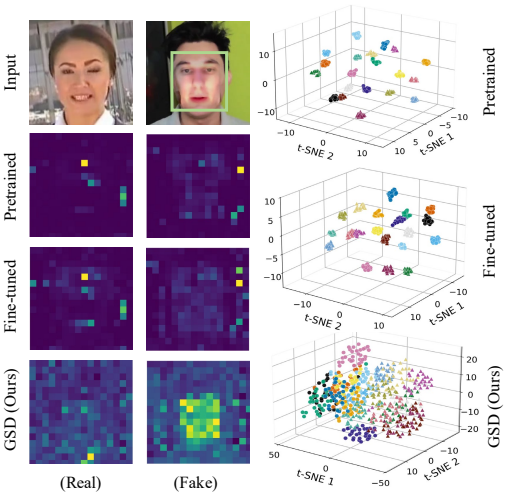


Figure 1: Attention and t-SNE (Maaten & Hinton, 2008) visualizations of pretrained CLIP, fine-tuned CLIP, and GSD-CLIP. In the t-SNE plots, colors denote **identities**, showing persistent identity-based semantic clustering after fine-tuning and its reduction by GSD. In the attention maps, naive fine-tuning retains patterns similar to pretrained CLIP, whereas GSD *redirects attention toward manipulated regions*.

Motivated by these observations, we propose a **Geometric Semantic Decoupling (GSD)** framework, which explicitly reduces the influence of semantic priors inherited from pretrained vision foundation models. Specifically, we use a frozen pretrained encoder as an anchor to characterize its dominant representation geometry via Singular Value Decomposition (SVD) (Golub & Van Loan, 2013), and suppress components of the trainable forensic representation aligned with this geometry. We further insert GSD into the last four transformer layers and design an adaptive suppression strategy that adjusts the layer-wise suppression strength according to the occupancy of the current forensic feature in the pretrained dominant subspace. To improve efficiency, we introduce a mini-batch-based approximation, termed *Batch-SVD*, which replaces repeated sample-wise decompositions with a shared subspace estimate and reduces the computational cost by more than $15\times$. Finally, to support both batch evaluation and realistic streaming scenarios, we develop three inference protocols, namely batch inference, per-sample inference, and reference-based inference.

Interestingly, after suppressing high-level semantic components, a clear structural change emerges in the feature space. Forgery-related cues become more prominent and shape the learned representation, as reflected by the greater dispersion of fake samples in the t-SNE visualization (Figure 1). A similar shift is observed in the attention maps: the naively fine-tuned detector retains an attention pattern similar to that of the pretrained CLIP encoder, whereas GSD substantially alters this behavior. Specifically, *attention on fake images becomes concentrated on manipulated regions despite the absence of explicit forgery-region supervision*, while attention on real images remains more diffuse because no localized manipulation cues are present. In summary, our main contributions are:

- We systematically characterize *semantic fallback* in VFM-based detectors, showing that semantic priors inherited from pretraining remain dominant in the representation space after forensic fine-tuning and are closely associated with cross-domain detection failures.
- We introduce **Geometric Semantic Decoupling (GSD)**, which uses a frozen CLIP encoder to estimate the dominant pretrained subspace and geometrically suppress its influence on the trainable representation. GSD further employs a layer-aware, sample-adaptive suppression coefficient λ, modulated by pretrained-geometry occupancy, to balance prior suppression and forensic preservation.
- We introduce a mini-batch SVD approximation, termed *Batch-SVD*, to exploit the structured and stable semantic subspace of CLIP, replacing costly per-sample subspace estimation and reducing the SVD cost by over $15\times$ in both training and inference. We further develop three complementary inference protocols (batch, per-sample, and reference-based inference) to support both offline batch evaluation and online streaming detection.

2 SEMANTIC FALLBACK AS A BOTTLENECK FOR FORGERY DETECTION

2.1 SEMANTIC FALLBACK IN FINE-TUNED REPRESENTATIONS

Table 1: Distribution-level identity similarity gap ΔS across CLIP layers.

Layer L	**2**	**4**	**8**	**12**	**24**
FaceForensics++					
$CLIP_{fz}$	0.153	0.102	0.153	0.145	0.333
$CLIP_{ft}$	0.152	0.110	0.142	0.136	0.317
Celeb-DF v2					
$CLIP_{fz}$	0.113	0.078	0.130	0.123	0.385
$CLIP_{ft}$	0.111	0.085	0.127	0.115	0.398
DFDC					
$CLIP_{fz}$	0.176	0.156	0.204	0.164	0.392
$CLIP_{ft}$	0.173	0.167	0.199	0.156	0.413

We begin by asking whether forensic fine-tuning truly removes the semantic priors inherited from pretraining. To this end, we use identity structure as an observable probe of high-level semantic organization in the representation space.

Definition 2.1 (Identity Similarity Gap)**.** *Given the features* $\{\boldsymbol{f}_i\}_{i=1}^N$ *with identity labels* $\{y_i\}_{i=1}^N$*, where* $\boldsymbol{f}_i \in \mathbb{R}^d$ *and* $y_i \in \{1, \dots, C\}$*, the identity similarity gap is defined as* $\Delta S \triangleq S_{\text{intra}} - S_{\text{inter}}$*, where* $S_{\text{intra}} \triangleq \mathbb{E}\left[\cos(\boldsymbol{f}_i, \boldsymbol{f}_j) \mid y_i = y_j\right]$ *and* $S_{\text{inter}} \triangleq \mathbb{E}\left[\cos(\boldsymbol{f}_i, \boldsymbol{f}_j) \mid y_i \neq y_j\right]$*.*

Using the identity similarity gap ΔS defined above, we evaluate both the frozen CLIP encoder (CLIP_{fz}) and its fine-tuned counterpart (CLIP_{ft}). Specifically, we extract features from layers $L \in \{2, 4, 8, 12, 24\}$ (with layer 24 denoting the final layer) and compute ΔS on the in-domain FaceForensics++ dataset and two cross-domain benchmarks (Table 1). We find that: 1) representations from every layer of both CLIP_{fz} and CLIP_{ft} exhibit clear identity-based organization across all datasets, indicating that forensic fine-tuning does not eliminate the high-level semantic priors inherited from pretraining. 2) The final-layer behavior differs between in-domain and cross-domain evaluation. On FaceForensics++ (Rossler et al., 2019), the final-layer ΔS decreases after fine-tuning, suggesting that the influence of inherited semantic priors is partially weakened when familiar manipulation cues are available. By contrast, on Celeb-DF v2 (Li et al., 2020b) and DFDC (Dolhansky et al., 2020), the final-layer ΔS is maintained or further increased after fine-tuning (from 0.385 to 0.398 and from 0.392 to 0.413), indicating that inherited semantic priors remain strongly expressed in the representation space on unseen domains.

2.2 HOW SEMANTIC FALLBACK RELATES TO PREDICTION RESULTS

We next examine whether stronger reliance on inherited semantic priors is associated with prediction failures using a sample-level metric termed *Pretrained Geometry Occupancy*.

Definition 2.2 (Pretrained Geometry Occupancy)**.** *For the* i*-th sample, let* $\bar{\boldsymbol{P}}_i \in \mathbb{R}^{N \times d}$ *denote the non-CLS patch-token representations extracted from the frozen CLIP encoder, where* N *is the number of patch tokens and* d *is the feature dimension. We perform singular value decomposition* $\bar{\boldsymbol{P}}_i = \boldsymbol{Q}_i \boldsymbol{\Sigma}_i \boldsymbol{V}_i^\top$ *and use the top-*K *right singular vectors* $\boldsymbol{V}_i^{(K)} \in \mathbb{R}^{d \times K}$ *to characterize the dominant subspace of the pretrained representation. Given the corresponding fine-tuned feature* $\boldsymbol{f}_i$*, its occupancy is defined as* $r_i = \left\| \boldsymbol{f}_i \boldsymbol{V}_i^{(K)} (\boldsymbol{V}_i^{(K)})^\top \right\|_F^2 / \left\| \boldsymbol{f}_i \right\|_F^2$.

We evaluate pretrained geometry occupancy on the FaceForensics++ and Celeb-DF v2 datasets. Given the substantial imbalance between correctly and incorrectly classified samples, we randomly sample balanced subsets with equal numbers of correct and incorrect samples and matched real/fake ratios. As shown in Figure 2, incorrectly classified samples are more concentrated in the high-occupancy region on both datasets, exhibiting higher minimum and mean occupancy than correctly classified samples. More notably, the overall occupancy distribution shifts substantially upward on Celeb-DF v2, with even correctly classified cross-domain samples exhibiting considerably higher occupancy than their in-domain counterparts. These results provide sample-level evidence for semantic fallback: domain shift increases the overall reliance on inherited pretrained structure, while misclassification tends to be associated with stronger alignment to the pretrained dominant subspace.

2.3 HOW SUPPRESSING PRETRAINED COMPONENTS AFFECTS DETECTION

If semantic fallback is a key factor limiting detection generalization, then suppressing feature components associated with inherited semantic priors should improve cross-domain performance. We therefore estimate its pretrained dominant subspace from the frozen CLIP patch-token representations using SVD, and transform the fine-tuned feature as

$$\boldsymbol{f}_i^{(\lambda)} = \boldsymbol{f}_i - \lambda \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{f}_i,$$

where $\lambda \in [0, 1]$ controls the strength of pretrained-prior suppression. We train a linear probe using the transformed features on FaceForensics++ and test on Celeb-DF v2 (CDv2), DFDCP (Dolhansky et al., 2019) and DFDC. As shown in Table 2, suppressing the pretrained dominant components consistently improves cross-domain detection while leaving the in-domain performance on FaceForensics++ largely unchanged. The gains generally increase with the suppression strength. This benefit on unseen domains is consistent with the occupancy analysis above, which shows stronger coupling to the pretrained dominant geometry on cross-domain samples. These results support semantic fallback as an important factor underlying detection failures, particularly under domain shift.

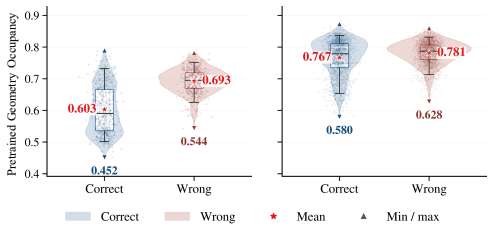


Figure 2: Pretrained geometry occupancy (r_i) of the fine-tuned CLIP on FaceForensics++ (left) and Celeb-DF v2 (right). Wrongly classified samples exhibit higher pretrained geometry occupancy statistics.

Table 2: Video-level detection AUC (%) under different pretrained-prior suppression strengths.

Dataset	λ=0	λ=0.2	λ=0.4	λ=0.6	λ=0.8
FF++	85.3	85.6	**85.7**	85.6	85.5
CDv2	79.1	79.6	79.8	79.9	**80.6**
DFDCP	63.8	64.0	64.2	64.4	**65.9**
DFDC	71.9	72.4	72.9	73.1	**73.7**

2.4 Target-Risk Trade-off of Pretrained-Prior Suppression

The results above reveal a trade-off between suppressing inherited semantic priors and preserving task-relevant forensic evidence, suggesting that the two are not fully separable in the representation space. To provide an analytical interpretation of this trade-off, let $\mathcal{R}_{\text{tgt}}(\lambda)$ denote the target-domain risk under suppression strength λ. We model it using the local second-order approximation:

$$\mathcal{R}_{\text{tgt}}(\lambda) \approx \mathcal{R}_0 + (1-\lambda)^2\Gamma_{\text{sem}} + \lambda^2\Gamma_{\text{share}} + \lambda^2\Gamma_{\text{mis}}, \tag{1}$$

where $\Gamma_{\text{sem}} > 0$ measures the penalty from retaining domain-sensitive semantic priors, $\Gamma_{\text{share}} > 0$ captures the cost of suppressing shared task-relevant evidence, and $\Gamma_{\text{mis}} \geq 0$ accounts for estimation error in the dominant pretrained subspace. A detailed representation-level derivation is provided in Appendix A. Under equation 1, the approximate target risk is minimized at $\lambda^\star = \frac{\Gamma_{\text{sem}}}{\Gamma_{\text{sem}}+\Gamma_{\text{share}}+\Gamma_{\text{mis}}}$, which satisfies $\lambda^\star \in (0, 1)$ when $\Gamma_{\text{sem}} > 0$ and $\Gamma_{\text{share}} + \Gamma_{\text{mis}} > 0$. This result formalizes the empirical observation that semantic suppression should be neither absent nor excessive: moderate suppression reduces semantic interference while preserving useful forensic evidence.

3 Methodology

3.1 Geometric Pretrained-Prior Decoupling with Single-SVD

Motivated by the *semantic fallback* phenomenon analyzed in Section 2, we introduce a **Geometric Semantic Decoupling (GSD)** module to suppress pretrained semantic priors that interfere with transferable forensic discrimination. Our detector consists of a trainable forensic branch and a frozen anchor branch, both initialized from the same pretrained CLIP weights. The frozen branch preserves the semantic priors inherited from pretraining, including identity, object category, and scene semantics, while the trainable branch learns features for real/fake discrimination. At selected transformer layers, GSD estimates the dominant pretrained geometry from the frozen branch and suppresses its influence on the trainable features, yielding a geometrically decoupled representation for downstream detection. In this way, GSD reduces the interference of inherited semantic priors while preserving the trainable branch's capacity to learn transferable forensic evidence.

Specifically, for the i-th sample at transformer layer l, let $\boldsymbol{f}_{i,l} \in \mathbb{R}^{(N+1)\times d}$ and $\boldsymbol{g}_{i,l} \in \mathbb{R}^{(N+1)\times d}$ denote the token representations from the trainable forensic branch and the frozen CLIP anchor,

respectively, where N is the number of patch tokens and d is the hidden dimension. We denote the non-CLS token matrix of the frozen branch as $\bar{\boldsymbol{P}}_{i,l} \in \mathbb{R}^{N\times d}$ and perform singular value decomposition $\bar{\boldsymbol{P}}_{i,l} = \boldsymbol{U}_{i,l}\boldsymbol{\Sigma}_{i,l}\boldsymbol{V}_{i,l}^{\top}$. The top-$k$ right singular vectors, denoted by $\boldsymbol{V}_{i,l}^{(k)}$, capture the dominant directions of variation in the frozen representation.

To determine the suppression strength, we build on the observation in Section 2 that misclassified samples exhibit higher pretrained geometry occupancy, suggesting that stronger alignment with the frozen pretrained geometry is associated with greater detection difficulty. We therefore quantify this alignment using pretrained geometry occupancy. Specifically, given the dominant pretrained basis $\boldsymbol{V}_{i,l}^{(k)}$, we define the pretrained geometry occupancy of the i-th sample at layer l as

$$r_{i,l} = \frac{|\overbrace{\boldsymbol{f}_{i,l}\boldsymbol{V}_{i,l}^{(k)}(\boldsymbol{V}_{i,l}^{(k)})^{\top}}^{\boldsymbol{f}_{i,l,\|}}|_F^2}{|\boldsymbol{f}_{i,l}|_F^2}, \tag{2}$$

where $\boldsymbol{f}_{i,l,\|}$ denotes the prior-aligned component obtained by projecting $\boldsymbol{f}_{i,l}$ onto the pretrained dominant subspace estimated from the frozen CLIP anchor. A larger $r_{i,l}$ indicates stronger alignment with the pretrained geometry. Since pretrained geometry occupancy varies across samples and the degree of alignment also differs across layers, as shown in Section 2, we adopt a layer-aware and sample-adaptive suppression coefficient as defined in equation 3. Here, $\sigma(\cdot)$ is the sigmoid function, $\lambda_{\max}$ denotes the maximum suppression strength, and a_l and b_l are learnable parameters for layer l.

$$\hat{\boldsymbol{f}}_{i,l} = \boldsymbol{f}_{i,l} - \underbrace{\lambda_{\max}\cdot\sigma(a_l r_{i,l} + b_l)}_{\lambda_{i,l}}\boldsymbol{f}_{i,l,\|} \tag{3}$$

After applying GSD at the selected transformer layers, the detector is optimized end-to-end with equation 4, where $\hat{\boldsymbol{f}}_i$ denotes the representation of the i-th sample, $t_i \in \{0,1\}$ is its ground-truth label, and $\boldsymbol{w}_c$ and b_c are the classifier weight and bias for class c, respectively.

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\boldsymbol{w}_{t_i}^{\top}\hat{\boldsymbol{f}}_i + b_{t_i}\right)}{\sum_{c=0}^{1}\exp\left(\boldsymbol{w}_c^{\top}\hat{\boldsymbol{f}}_i + b_c\right)} \tag{4}$$

3.2 Batch-SVD Approximation for Efficient Training

Applying a separate SVD to every sample at each selected transformer layer is computationally expensive, especially for large backbones such as ViT-L/14. As shown in Table 3, SVD alone accounts for over 88% of the total training time and over 93% of the total inference time under Single-SVD. We therefore introduce a Batch-SVD approximation that replaces repeated per-sample decompositions with a single decomposition over the aggregated frozen representations, amortizing the subspace-estimation cost across the mini-batch. Specifically, for the i-th sample, let $\bar{\boldsymbol{P}}_{i,l} \in \mathbb{R}^{N\times d}$ denote the non-CLS token matrix extracted from the frozen semantic anchor. We concatenate these matrices along the token dimension and perform a single singular value decomposition:

$$\bar{\boldsymbol{M}}_l = \text{Concat}_b\left(\bar{\boldsymbol{P}}_{1,l}, \bar{\boldsymbol{P}}_{2,l}, \ldots, \bar{\boldsymbol{P}}_{B,l}\right) = \boldsymbol{U}_l^{(b)}\boldsymbol{\Sigma}_l^{(b)}(\boldsymbol{V}_l^{(b)})^{\top}, \quad \bar{\boldsymbol{M}}_l \in \mathbb{R}^{BN\times d}. \tag{5}$$

We use the top-k right singular vectors of $\boldsymbol{V}_l^{(b)}$ to form a shared basis $\boldsymbol{V}_l^{(b,k)} \in \mathbb{R}^{d\times k}$, which replaces the sample-specific basis in equation 2 to compute the decoupled representation $\tilde{\boldsymbol{f}}_{i,l}$.

Batch-SVD substantially reduces the computational cost of pretrained subspace estimation. Under a setting of 10,000 images with a batch size of 128 on a single NVIDIA A100 (80GB) GPU, Batch-SVD reduces the per-image SVD cost by more than $15\times$ compared with Single-SVD in both training and inference. As shown in Table 3, the share of total runtime spent on SVD decreases from 88.42% to 32.72% during training and from 93.90% to 50.45% during inference.

3.3 Inference Protocols in Real-world Detection

We consider three protocols to support both *offline batch evaluation* and *online streaming inference*.

Table 4: Our method outperforms existing methods on five cross-domain datasets. "–" indicates that the results are not reported in the original paper, and † indicates that the results are obtained by using the official pretrained model or reproduction, and some results are cited from (Yan et al., 2024b; Chen et al., 2026). Top results are highlighted in **bold**.

Methods	Video-level AUC						Methods	Frame-level AUC					
	CDv2	CDv1	DFDC	DFDCP	DFD	**Avg.**		CDv2	CDv1	DFDC	DFDCP	DFD	**Avg.**
F3Net (Qian et al., 2020) (ECCV)	78.9	–	71.8	74.9	84.4	77.5	Efficient-b4† (Tan & Le, 2019) (ICML)	74.9	70.5	69.6	72.8	81.5	73.9
CLIP† (Radford et al., 2021) (ICML)	88.4	89.6	83.3	86.3	92.6	88.0	CLIP† (Radford et al., 2021) (ICML)	83.2	86.1	80.7	82.4	87.7	84.0
LaV (Shuai et al., 2023) (MM)	92.2	84.7	82.0	89.7	96.8	89.1	Face X-Ray (Li et al., 2020a) (CVPR)	67.9	70.9	63.3	69.4	76.7	69.6
IID (Huang et al., 2023) (CVPR)	83.8	–	70.0	68.9	93.9	79.2	RECCE (Cao et al., 2022) (CVPR)	73.2	–	71.3	74.2	74.2	73.2
CFM (Luo et al., 2023) (TIFS)	89.7	–	70.6	80.2	95.2	83.9	UCF† (Yan et al., 2023a) (ICCV)	82.4	77.9	80.5	75.9	87.8	80.9
SBIs (Shiohara & Yamasaki, 2022) (CVPR)	90.6	–	75.2	87.7	88.2	85.4	ICT (Dong et al., 2022) (CVPR)	85.7	81.4	–	–	84.1	83.7
ProDet† (Cheng et al., 2024) (NeurIPS)	92.6	95.7	70.7	82.8	95.4	87.4	ProDet† (Cheng et al., 2024) (NeurIPS)	84.5	90.9	72.4	81.2	–	82.3
CDFA† (Lin et al., 2024) (ECCV)	93.8	92.1	83.0	88.1	95.4	90.5	DFFake (Sun et al., 2024) (NeurIPS)	80.5	–	–	81.0	90.4	84.0
FakeRadar (Li et al., 2025) (ICCV)	91.7	–	84.1	88.5	96.2	90.1	ExposDe† (Ba et al., 2024) (AAAI)	86.4	81.8	72.1	85.1	89.8	83.0
VbSaT (Yan et al., 2025) (CVPR)	94.7	–	84.3	90.9	96.5	91.6	LSDA (Yan et al., 2024a) (CVPR)	83.0	86.7	73.6	81.5	88.0	82.6
DeepShield (Cai et al., 2025) (ICCV)	92.2	–	82.8	**93.2**	96.1	91.1	UDD (Fu et al., 2025) (AAAI)	86.9	–	75.8	85.6	91.0	84.8
FCG (Han et al., 2025) (CVPR)	95.0	–	81.8	–	–	88.4	FSgD (Ma et al., 2025) (NeurIPS)	86.7	90.1	–	81.8	82.1	85.2
Effort† (Yan et al., 2024b) (ICML)	95.6	91.8	84.3	90.9	96.5	91.8	FreqDebias (Kashiani et al., 2025) (CVPR)	83.6	87.5	74.1	82.4	86.8	82.9
χ^2-DFD (Chen et al., 2026) (NeurIPS)	95.5	–	85.3	91.2	95.7	91.9	RAID (Cheng et al., 2025) (NeurIPS)	84.4	–	66.9	–	91.2	80.8
GSD-CLIP	**96.4**	**97.0**	**88.3**	**93.2**	**97.7**	**94.5**	GSD-CLIP	**90.6**	**91.3**	**85.7**	**90.1**	**94.5**	**90.4**

Batch Inference. For standard benchmark evaluation, we directly use the Batch-SVD formulation adopted during training. At each selected layer, the frozen anchor estimates a shared pretrained basis from the current mini-batch, and each sample is decoupled using its adaptive suppression coefficient. This protocol is efficient for large-scale dataset-level evaluation.

Per-sample Inference. For online or streaming detection, where inputs arrive independently, we use the Single-SVD formulation in Section 3.1. The frozen anchor estimates a sample-specific pretrained subspace for each test image, and the trainable representation is decoupled accordingly. This protocol is suitable for real-world streaming scenarios but incurs a higher computational cost.

Table 3: Per-image runtime comparison of Batch-SVD and Single-SVD. The $ratio$ denotes the percentage of total computation time spent on SVD operations.

Setting	SVD Time (ms)	Ratio (%)	Speedup
Training (per image)			
Batch-SVD	5.43 ± 0.02	32.72 ± 0.09	$\sim 15.6\times$
Single-SVD	84.81 ± 0.02	88.42 ± 0.07	$1\times$
Inference (per image)			
Batch-SVD	3.37 ± 0.02	50.45 ± 0.05	$\sim 15.2\times$
Single-SVD	51.13 ± 0.03	93.90 ± 0.06	$1\times$

Reference-based Inference. We further introduce reference-based inference, which estimates a shared pretrained subspace from a reference set of frozen CLIP features. The reference images only need to share the same high-level semantic category and need not come from the test set, enabling efficient inference without access to target-domain data. We further provide evidence for the compatibility of these inference protocols with Batch-SVD-trained detectors in Appendix B.

4 Experiments

We adopt the pretrained CLIP ViT-L/14 (Radford et al., 2021) as the default backbone and optimize the entire model end-to-end using AdamW (Loshchilov & Hutter, 2017). The backbone learning rate is set to 1×10^{-6}, with batch sizes of 128 for training and 512 for testing. All experiments are conducted on NVIDIA A100 GPUs. Following (Yan et al., 2024a; 2023b), we apply standard augmentations, including Gaussian blur and JPEG compression. GSD is applied to the last four CLIP transformer layers by default, with learnable modulation parameters initialized as $a = 2$ and $b = 1$, $\lambda_{\max} = 1$, and $k = 64$ unless otherwise specified; batch inference is used by default.

4.1 Face Forgery Detection

We train all models exclusively on FaceForensics++ (Rossler et al., 2019) following DeepfakeBench (Yan et al., 2023b). Under the **cross-dataset** setting, as shown in Table 4, GSD achieves the best overall performance at both video and frame levels, reaching an average video-level AUC of 94.5%, outperforming χ^2-DFD (Chen et al., 2026) (91.9%) by 2.6%, and an average frame-level AUC of 90.4%, surpassing FSgD (Ma et al., 2025) (85.2%) by 5.2%. Under the **cross-manipulation** setting, as shown in Table 5, GSD also achieves the best overall performance, with average video- and frame-level AUCs of 97.8% and 94.6%, respectively. These results indicate that reducing inter-

Table 5: Our method achieves superior results on unseen manipulation methods. Results are reported on six representative face-swapping methods selected from DF40.

Methods	Video-level AUC							Methods	Frame-level AUC						
	UniFace	e4s	FaceDancer	FSGAN	InSwap	SimSwap	**Avg.**		UniFace	e4s	FaceDancer	FSGAN	InSwap	SimSwap	**Avg.**
CLIP (Radford et al., 2021) (ICML)	91.0	91.6	89.8	96.0	88.5	80.3	89.5	CLIP (Radford et al., 2021) (ICML)	86.5	87.8	85.6	92.0	83.9	76.3	85.4
RECCE (Cao et al., 2022) (CVPR)	89.8	68.3	84.8	94.9	84.8	76.8	83.2	RECCE (Cao et al., 2022) (CVPR)	84.2	65.2	78.3	88.4	79.5	73.0	78.1
SBI (Shiohara & Yamasaki, 2022) (CVPR)	72.4	75.0	59.4	80.3	71.2	70.1	71.4	IID (Huang et al., 2023) (CVPR)	79.5	71.0	79.0	86.4	74.4	64.0	75.7
UCF (Yan et al., 2023a) (ICCV)	83.1	73.1	86.2	93.7	80.9	64.7	80.3	SBI (Shiohara & Yamasaki, 2022) (CVPR)	64.4	69.0	44.7	87.9	63.3	56.8	64.4
IID (Huang et al., 2023) (CVPR)	83.9	76.6	84.4	92.7	78.9	64.4	80.2	UCF (Yan et al., 2023a) (ICCV)	78.7	69.2	80.0	88.1	76.8	64.9	76.3
LSDA (Yan et al., 2024a) (CVPR)	87.2	69.4	72.1	93.9	85.5	69.3	79.6	LSDA (Yan et al., 2024a) (CVPR)	85.4	68.4	75.9	83.2	81.0	72.7	77.8
ProDet (Cheng et al., 2024) (NeurIPS)	90.8	77.1	74.7	92.8	83.7	84.4	83.9	CDFA (Lin et al., 2024) (ECCV)	76.5	67.4	75.4	84.8	72.0	76.1	75.4
CDFA (Lin et al., 2024) (ECCV)	76.2	63.1	80.3	94.2	77.2	75.7	77.8	ProDet (Cheng et al., 2024) (NeurIPS)	84.5	71.0	73.6	86.5	78.8	77.8	78.7
Effort (Yan et al., 2024b) (ICML)	96.2	98.3	92.6	95.7	93.6	92.6	94.8	FSgD (Ma et al., 2025) (NeurIPS)	91.8	87.5	83.0	86.3	87.4	91.0	87.8
VbSaT (Yan et al., 2025) (CVPR)	96.0	98.0	91.6	96.4	93.7	93.1	94.8	χ^2-DFD (Chen et al., 2026) (NeurIPS)	85.2	91.2	83.8	89.9	78.4	84.9	85.6
GSD-CLIP	**98.6**	**99.5**	**97.5**	**98.8**	**96.5**	**96.0**	**97.8**	GSD-CLIP	**96.3**	**98.0**	**93.8**	**96.1**	**92.1**	**91.1**	**94.6**

Table 6: ACC (%) performance on the UniversalFakeDetect dataset. Our method achieves the best average performance and demonstrates competitive results across most generative categories.

Methods	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	**Avg.**
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNN-Spot (Wang et al., 2020) (CVPR)	**100.0**	85.2	70.2	85.7	79.0	91.7	53.5	66.7	48.7	86.3	86.3	60.1	54.0	55.0	54.1	60.8	63.8	65.7	55.6	69.6
Patchfor (Chai et al., 2020) (ECCV)	75.0	69.0	68.5	79.2	64.2	63.9	75.5	75.1	75.3	72.3	55.3	67.4	76.5	76.1	75.8	74.8	73.3	68.5	67.9	71.2
Freq-spec (Zhang et al., 2019) (WIFS)	49.9	**99.9**	50.5	49.9	50.3	99.7	50.1	50.0	48.0	50.6	50.1	50.9	50.4	50.4	50.3	51.7	51.4	50.4	50.0	55.5
F3Net (Qian et al., 2020) (ECCV)	99.4	76.4	65.3	92.6	58.1	**100.0**	63.5	54.2	47.3	51.5	51.5	69.2	68.2	75.3	68.8	81.7	83.2	83.0	66.3	71.3
CLIP† (Radford et al., 2021) (ICML)	**100.0**	97.0	98.1	82.3	**100.0**	87.6	53.0	59.1	50.0	73.1	90.3	48.0	81.3	61.6	80.4	57.2	59.8	57.9	81.4	74.6
UniFD (Ojha et al., 2023) (CVPR)	**100.0**	98.5	94.5	82.0	99.5	97.0	66.6	63.0	57.5	59.5	72.0	70.0	94.2	73.8	94.4	79.1	79.8	78.1	86.8	81.4
LGrad (Tan et al., 2023) (CVPR)	99.8	85.4	82.9	94.8	72.5	99.6	58.0	62.5	50.0	50.7	50.8	77.5	94.2	95.8	94.8	87.4	90.7	89.5	88.3	80.3
FreqNet (Tan et al., 2024a) (AAAI)	97.9	95.8	90.5	97.5	90.2	93.4	97.4	88.9	59.0	71.9	67.3	86.7	84.5	**99.6**	65.6	85.7	97.4	88.2	59.1	85.1
NPR (Tan et al., 2024b) (CVPR)	99.8	95.0	87.5	96.2	86.6	99.8	76.9	66.9	**98.6**	50.0	50.0	84.5	97.7	98.0	98.2	96.2	97.2	97.3	87.2	87.6
FatFormer (Liu et al., 2024a) (CVPR)	99.9	99.3	99.5	97.2	99.4	99.8	93.2	81.1	68.0	69.5	69.5	76.0	98.6	94.9	98.7	94.3	94.7	94.2	98.8	90.9
C2P-CLIP (Tan et al., 2025) (AAAI)	**100.0**	97.3	99.1	96.4	99.2	99.6	93.8	**95.6**	64.4	93.3	93.3	69.1	99.3	97.3	99.3	95.3	95.3	96.1	98.6	93.8
MPFT (Niu et al., 2026) (Arxiv)	99.9	99.2	97.3	94.9	97.6	99.7	**97.6**	94.7	90.9	87.7	84.8	82.0	99.3	97.2	99.5	95.5	96.3	96.8	98.4	95.2
PPL (Yang et al., 2026) (ICLR)	89.12	97.16	83.57	89.94	75.29	97.12	–	–	–	–	–	**94.55**	98.80	98.80	98.80	96.90	97.40	**97.65**	98.80	93.9
GSD-CLIP	**100.0**	99.4	**99.7**	**98.3**	99.8	**100.0**	91.4	90.5	91.1	**98.3**	**98.3**	91.9	**99.6**	98.2	**99.7**	**97.8**	**98.4**	97.2	**99.5**	**97.3**

ference from inherited semantic priors enables the detector to focus more on forgery-related evidence and achieve better generalization to unseen forgery pipelines.

4.2 Synthetic Image Detection

Following the protocol in (Yan et al., 2024b), we train the detector only on ProGAN, and evaluate cross-generator generalization on UniversalFakeDetect. As shown in Table 6, our method achieves the best average accuracy of **97.3%**, outperforming MPFT (Niu et al., 2026) (95.2%) by **2.1%** and C2P-CLIP (Tan et al., 2025) (93.8%) by **3.5%**. The gains are broadly distributed across diverse generator families, including challenging unseen regimes such as perceptual-loss-based generation, Glide, and DALL-E. Although our method is not superior on every individual subset, it achieves the highest overall average and more balanced performance across heterogeneous generators.

4.3 Comparison of Training and Inference Protocol

We first examine different training and inference protocols in the face forgery setting. The detector is evaluated on 11 cross-domain test sets. For Reference-based Inference, we construct two 512-image reference sets: one from FF++ and another from CDv2 to simulate an external distribution. As shown in Table 7, models trained with Batch-SVD achieve similar performance under Batch, Per-sample, and Reference-based Inference. Reference-based Inference also remains competitive with CDv2 references, indicating that it does not require samples from the training distribution. In addition, Single-SVD training provides only marginal gains over Batch-SVD, suggesting that mini-batch-level estimation is sufficient to capture the pretrained geometry of the frozen CLIP anchor.

We further evaluate inference strategies in the natural-image setting using the CycleGAN and ProGAN subsets of UniversalFakeDetect, where explicit semantic categories enable category-aligned reference construction. We sample matched natural images from ImageNet-12K and Pascal VOC 2012 as external references. As reported in Table 8, Per-sample and Batch Inference achieve nearly identical performance, and Reference-based Inference remains effective with out-of-distribution reference images. These results show that GSD does not require the training-time subspace estimation protocol to be exactly reproduced at test time. Once the decoupled representation is learned, different inference strategies remain well anchor-aligned, allowing efficient Batch-SVD training while supporting both streaming and reference-based inference.

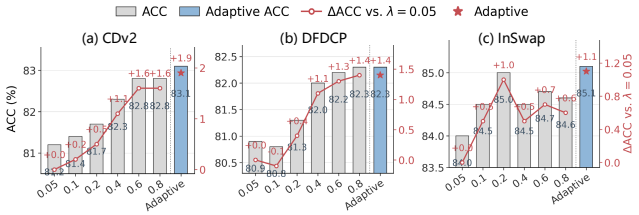


Figure 3: Comparison under different subtraction strategies. We report ACC (%) on the testing datasets.

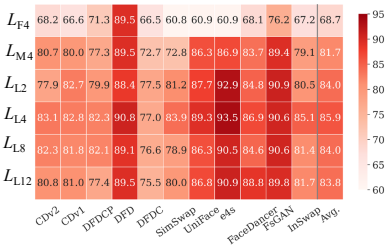


Figure 4: Comparison (ACC%) of six insertion strategies.

4.4 Ablation Study

The Impact of Different λ. We investigate the effect of the suppression coefficient λ by replacing the adaptive coefficient with a set of fixed constants. The results are reported in Figure 3. We observe that when λ is fixed, ACC increases steadily as λ grows from 0.05 to 0.6, and then tends to saturate at larger values. However, on InSwap, the best performance is achieved at $\lambda = 0.2$, indicating that the optimal suppression strength varies across datasets. In this case, the adaptive λ consistently outperforms all fixed settings, suggesting that sample-dependent suppression better balances pretrained prior interference and discriminative forensic cues.

The Impact of the GSD Module Placement. We evaluate six insertion depths: the first (L_{F4}) and middle (L_{M4}) four layers, as well as the last 2 (L_{L2}), 4 (L_{L4}), 8 (L_{L8}), and 12 (L_{L12}) layers. The results in Figure 4 show that GSD should not be inserted too early. Applying it to the first four layers leads to a substantial performance drop, while inserting it into the middle four layers or the last 8/12 layers remains clearly inferior to inserting it into the last four layers. This is because shallow layers retain more low-level texture and artifact cues, whereas deeper layers exhibit a more coherent dominant pretrained geometry; this observation is consistent with prior findings that preserving shallow forensic details is important for manipulation detection (Shuai et al., 2023; Nguyen et al., 2024). At the same time, the insertion depth should not be excessively late. For example, inserting GSD only into the last two layers performs noticeably worse than inserting it into the last four layers, suggesting that moderately deep layers provide a more suitable representation space for semantic-prior suppression.

Table 7: Comparison of different training and inference protocols for face forgery detection. ACC and frame-level AUC (%) are reported on representative datasets, along with the average results over all 11 test datasets.

Infer	Ref. Dataset	Metric	CDv2	DFDC	e4s	FaceDancer	Avg.all
Single-SVD Training for Face Forgery							
P-Infer	–	ACC	83.0	**77.2**	**93.7**	**87.3**	**86.0**
P-Infer	–	AUC	90.4	**86.2**	**98.1**	**94.1**	**92.8**
Batch-SVD Training for Face Forgery							
P-Infer	–	ACC	82.7	76.4	93.1	87.0	85.6
P-Infer	–	AUC	90.4	85.5	97.9	93.9	92.4
R-Infer	FF++	ACC	82.6	76.9	93.3	86.7	85.7
R-Infer	FF++	AUC	90.3	85.5	97.9	93.7	92.3
R-Infer	CDv2	ACC	82.6	76.4	93.2	86.7	85.7
R-Infer	CDv2	AUC	90.4	85.5	97.9	93.7	92.3
B-Infer	–	ACC	**83.1**	77.0	93.5	86.9	85.9
B-Infer	–	AUC	**90.6**	85.8	98.0	93.8	92.7

Table 8: Comparison of different inference protocols for the natural-image setting.

Infer	Metric	CycleGAN				
		Apple	Horse	Orange	Summer	Avg.all
R-Infer	ACC	99.6	**99.8**	99.4	99.3	99.4
B-Infer	ACC	99.6	99.7	99.4	99.3	99.4
Infer	**Metric**	**ProGAN**				
		Airplane	**Bicycle**	**Car**	**Dog**	**Avg.all**
R-Infer	ACC	100.0	100.0	100.0	100.0	100.0
B-Infer	ACC	100.0	100.0	100.0	100.0	100.0

Compatibility with Various Vision Foundation Models. We further evaluate GSD with different vision foundation models, including DINOv2 (Oquab et al., 2023), BLIP (Li et al., 2022), SigLIP (Zhai et al., 2023), and CLIP. As shown in Table 9, GSD consistently improves the average AUC across all backbones. The gains are also consistent across evaluation datasets, indicating that GSD can be readily integrated into diverse VFMs to improve cross-domain generalization.

The Impact of Different Batch Sizes. We also evaluate the effects of both training and inference batch sizes, as reported in Table 10. Reducing the training batch size from 128 to 16 only slightly decreases the average frame-level AUC from 92.3% to 92.0%. The effect of inference batch size is even smaller: with the default training batch size, varying it from 32 to 256 produces almost no

performance change. Overall, GSD is robust to batch-size variation, enabling flexible deployment under limited GPU memory and computational resources.

5 Related Work

Generalizable AI-generated Image Detection Early AI-generated image detectors primarily relied on manipulation-specific artifacts, including frequency irregularities (Wang et al., 2020; Qian et al., 2020; Kashiani et al., 2025; Li et al., 2021), local texture patterns (Chai et al., 2020; Liu et al., 2021; Masi et al., 2020), gradient-domain inconsistencies (Tan et al., 2023), and face-specific spatial or temporal cues (Li et al., 2020a; Chen et al., 2022; Dang et al., 2020; Zhao et al., 2021; Haliassos et al., 2021; Dong et al., 2022; Cozzolino et al., 2021; Sun et al., 2022). Although effective in-domain, such supervised detectors often generalize poorly to unseen datasets or manipulation pipelines. More recently, a growing body of work (Ojha et al., 2023; Liu et al., 2024a; Tan et al., 2025; Liu et al., 2024b; Cui et al., 2025; Yan et al., 2024b; Han et al., 2025; Yan et al., 2025; Cai et al., 2025) has adopted Vision Foundation Models (VFMs), particularly CLIP (Radford et al., 2021), as strong backbones for AI-generated image detection due to their transferable visual representations. Despite improved performance, these methods largely focus on how to adapt or preserve pretrained representations, while paying less attention to how inherited semantic priors interact with forensic evidence under domain shift. Our work studies this issue explicitly and identifies *semantic fallback*, where inherited semantic priors remain influential after forensic fine-tuning and interfere with transferable forensic discrimination.

Table 9: Ablation studies (video-level AUC%) on different vision foundation models (VFMs).

Method	CDv2	DFDCP	DFDC	InSwap	FaceDancer	Avg.
BLIP	83.3	81.5	76.9	87.5	81.0	82.0
+Ours	87.4 +4.1	84.9 +3.4	80.3 +3.4	91.2 +3.7	85.7 +4.7	85.9 +3.9
SigLIP	88.0	86.3	78.7	85.8	82.5	84.3
+Ours	91.3 +3.3	89.4 +3.1	81.0 +2.3	93.1 +7.3	89.9 +7.4	88.9 +4.7
DINOv2	90.0	87.9	81.7	91.7	91.6	88.6
+Ours	93.4 +3.4	90.8 +2.9	85.6 +3.9	96.2 +4.5	95.0 +3.4	92.2 +3.6
CLIP	90.7	88.4	83.1	90.3	91.5	88.8
+Ours	96.4 +5.7	93.2 +4.8	88.6 +5.5	96.5 +6.2	97.5 +6.0	94.4 +5.6

Table 10: Effect of training and inference batch sizes on frame-level AUC (%).

Batchsize	CDv1	DFD	DFDC	FaceDancer	FSGAN	Avg.
Different Training Batch Sizes (Test batch size $4\times$)						
$B_{tr} = 16$	90.6	94.4	85.0	93.6	96.3	92.0
$B_{tr} = 32$	90.5	94.2	85.1	93.7	96.4	92.0
$B_{tr} = 64$	90.9	94.1	86.1	93.6	96.4	92.2
$B_{tr} = 128$	91.3	94.5	85.8	93.8	96.1	92.3
Different Inference Batch Sizes (Training Batch Size = 128)						
$B_{ts} = 32$	91.1	94.4	85.8	93.8	96.1	92.2
$B_{ts} = 64$	91.2	94.5	85.8	93.9	96.1	92.3
$B_{ts} = 128$	91.0	94.5	85.5	93.9	96.1	92.2
$B_{ts} = 256$	91.3	94.5	85.7	93.8	96.1	92.3

Disentanglement and Semantic Debiasing Some studies attempt to disentangle forgery-relevant cues from identity or other high-level attributes (Luo et al., 2023; Fu et al., 2025; Yan et al., 2023a; Sun et al., 2024; Ba et al., 2024). These methods typically rely on auxiliary reconstruction branches, token perturbation, or regularization objectives to reduce reliance on semantic features learned from the training data. Our work instead investigates how semantic priors already inherited from pretrained VFMs can limit detection generalization. EFFORT (Yan et al., 2024b) is particularly related because it also uses SVD (Golub & Van Loan, 2013) to structure VFM representations. However, EFFORT preserves the dominant pretrained subspace and adapts the residual components to retain useful pretrained knowledge. Our analysis, by contrast, shows that inherited semantic priors can become overly influential under domain shift and interfere with transferable forensic discrimination, motivating GSD to use the frozen pretrained representation as an anchor and explicitly suppress the corresponding components in the trainable forensic representation.

6 Conclusion and Limitations

In this paper, we identify *semantic fallback* as a key failure mode in VFM-based AI-generated image detectors, where semantic priors inherited from pre-training remain influential after fine-tuning and interfere with transferable forensic discrimination. To mitigate this effect, we propose Geometric Semantic Decoupling (GSD), a simple projection-based module that suppresses prior-aligned components in learned representations. GSD still has limitations. Although it adaptively suppresses pre-trained priors, it does not explicitly model which pre-trained components are beneficial or detrimental to transferable forensic discrimination. Consequently, it may not optimally balance prior suppression against forensic preservation. Future work could explicitly model forgery-specific evidence, such as frequency artifacts, to enable more selective suppression.

AI USE STATEMENT

We used generative AI tools to assist with writing, literature retrieval, and the formulation of mathematical arguments. All AI-assisted content, such as retrieved references, was manually reviewed and independently verified by the authors. The authors take full responsibility for all content, claims, results, and artifacts in this work.

REFERENCES

Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 719–728, 2024.

A. Brown. Algorithmic harms: The societal impact of non-consensual ai-generated content. *Nature Machine Intelligence (Perspective)*, 6:112–115, 2024.

CNN Business. Deepfake video conference call swindles multinational firm out of $25 million, 2024. URL https://www.cnn.com/2024/02/04/asia/deepfake-video-conference-scam-hong-kong-intl-hnk/index.html.

Yinqi Cai, Jichang Li, Zhaolun Li, Weikai Chen, Rushi Lan, Xi Xie, Xiaonan Luo, and Guanbin Li. Deepshield: Fortifying deepfake video detection with local and global forgery analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12524–12534, 2025.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4113–4122, 2022.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pp. 103–120. Springer, 2020.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18710–18719, 2022.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. *Advances in Neural Information Processing Systems*, 38:83792–83839, 2026.

Harry Cheng, Ming-Hui Liu, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Fair deepfake detectors can generalize. *arXiv preprint arXiv:2507.02645*, 2025.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37:21979–21998, 2024.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15108–15117, 2021.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19207–19217, 2025.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pp. 5781–5790, 2020.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9468–9478, 2022.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3040–3048, 2025.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5039–5049, 2021.

Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22995–23005, 2025.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4490–4499, 2023.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2898, 2020.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8775–8785. IEEE, 2025.

Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6458–6467, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020a.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2–5, 14, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020b.

Zhaolun Li, Jichang Li, Yinqi Cai, Junye Chen, Xiaonan Luo, Guanbin Li, and Rushi Lan. Fakeradar: Probing forgery outliers to detect unknown deepfake videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13382–13392, 2025.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European conference on computer vision*, pp. 104–122. Springer, 2024.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 772–781, 2021.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024a.

Zihan Liu, Hanyi Wang, Yaoyu Kang, and Shilin Wang. Mixture of low-rank experts for transferable ai-generated image detection. *arXiv preprint arXiv:2404.04883*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023.

Long Ma, Zhiyuan Yan, Jin Xu, Yize Chen, Qinglang Guo, Zhen Bi, Yong Liao, and Hui Lin. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *arXiv preprint arXiv:2504.04827*, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European conference on computer vision*, pp. 667–684. Springer, 2020.

Midjourney. Midjourney. `https://www.midjourney.com/`, 2022. Accessed: 2026-01-04.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17395–17405, 2024.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12–21, 2022.

Yakun Niu, Yingjian Chen, and Lei Zhang. Detecting ai-generated images via distributional deviations from real images. *arXiv preprint arXiv:2601.03586*, 2026.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18720–18729, 2022.

Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. Locate and verify: A two-stream network for improved deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7131–7142, 2023.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2316–2324, 2022.

Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22412–22423, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023b.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8984–8994, 2024a.

Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2, 2024b.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12615–12625, 2025.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, et al. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, volume 2026, pp. 121742–121761, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

A A REPRESENTATION VIEW OF SEMANTIC SUPPRESSION

We introduce a coupled-representation view to provide an analytical interpretation of how suppressing pretrained priors may affect generalization and why its strength requires careful control.

Assumption A.1 (Coupled Representation). *Let $\boldsymbol{f} \in \mathbb{R}^d$ denote the learned feature representation of an input sample. We assume that*

$$\boldsymbol{f} = \boldsymbol{f}_{\text{for}} + \boldsymbol{f}_{\text{sem}} + \boldsymbol{f}_{\text{shr}} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{f}_{\text{for}}$ denotes task-relevant forensic evidence that is comparatively stable across domains, $\boldsymbol{f}_{\text{sem}}$ denotes the semantic component inherited from pretrained representations that may interfere with forensic decisions and induce domain-sensitive shortcuts, $\boldsymbol{f}_{\text{shr}}$ denotes a shared component jointly associated with pretrained semantic priors and task-relevant forensic evidence, and $\boldsymbol{\varepsilon}$ collects residual variation not captured by the preceding structured terms.

Definition A.1 (Pretrained-Prior Suppression). *Let $\boldsymbol{P}_{\hat{\mathcal{P}}} = \boldsymbol{U}\boldsymbol{U}^\top$ denote the orthogonal projector onto an estimated pretraining-dominant subspace $\hat{\mathcal{P}}$, where $\boldsymbol{U}$ contains an orthonormal basis of $\hat{\mathcal{P}}$. For a suppression strength $\lambda \in [0, 1]$, define*

$$T_\lambda(\boldsymbol{f}) = \left(\boldsymbol{I} - \lambda \boldsymbol{P}_{\hat{\mathcal{P}}}\right) \boldsymbol{f}. \tag{6}$$

Here, $\hat{\mathcal{P}}$ captures dominant representation directions inherited from pretraining and may deviate from the ideal pretrained-prior subspace $\mathcal{P}^\star$ due to finite-sample estimation error, domain shift, or imperfect subspace estimation. Applying T_λ to the decomposition in Assumption A.1 yields

$$\begin{aligned} T_\lambda(\boldsymbol{f}) = &\left(\boldsymbol{I} - \lambda \boldsymbol{P}_{\hat{\mathcal{P}}}\right) \boldsymbol{f}_{\text{for}} + \left(\boldsymbol{I} - \lambda \boldsymbol{P}_{\hat{\mathcal{P}}}\right) \boldsymbol{f}_{\text{sem}} \\ &+ \left(\boldsymbol{I} - \lambda \boldsymbol{P}_{\hat{\mathcal{P}}}\right) \boldsymbol{f}_{\text{shr}} + \left(\boldsymbol{I} - \lambda \boldsymbol{P}_{\hat{\mathcal{P}}}\right) \boldsymbol{\varepsilon}. \end{aligned} \tag{7}$$

Proposition A.1 (Margin Shift under Pretrained-Prior Suppression). *Let $h(\boldsymbol{f}) = \boldsymbol{w}^\top \boldsymbol{f}$ be a linear classifier, and let $m_\lambda = y\, \boldsymbol{w}^\top T_\lambda(\boldsymbol{f})$ denote the signed margin under pretrained-prior suppression, where $y \in \{-1, +1\}$ is the ground-truth label. Under Assumption A.1,*

$$\begin{aligned} m_\lambda = &\, y\, \boldsymbol{w}^\top \left(\boldsymbol{f}_{\text{for}} + \boldsymbol{f}_{\text{sem}} + \boldsymbol{f}_{\text{shr}} + \boldsymbol{\varepsilon}\right) \\ &- \lambda y\, \boldsymbol{w}^\top \boldsymbol{P}_{\hat{\mathcal{P}}} \left(\boldsymbol{f}_{\text{for}} + \boldsymbol{f}_{\text{sem}} + \boldsymbol{f}_{\text{shr}} + \boldsymbol{\varepsilon}\right). \end{aligned} \tag{8}$$

Equivalently,

$$m_\lambda - m_0 = -\lambda y\, \boldsymbol{w}^\top \boldsymbol{P}_{\hat{\mathcal{P}}} \left(\boldsymbol{f}_{\text{for}} + \boldsymbol{f}_{\text{sem}} + \boldsymbol{f}_{\text{shr}} + \boldsymbol{\varepsilon}\right), \tag{9}$$

where $m_0 = y\, \boldsymbol{w}^\top \boldsymbol{f}$ is the unsuppressed margin.

This result follows directly from the linearity of T_λ. Proposition A.1 shows that pretrained-prior suppression not only attenuates the pretraining-dominant component $\boldsymbol{P}_{\hat{\mathcal{P}}} \boldsymbol{f}_{\text{sem}}$, but may also suppress the shared component $\boldsymbol{P}_{\hat{\mathcal{P}}} \boldsymbol{f}_{\text{shr}}$ and useful forensic evidence when $\hat{\mathcal{P}}$ is imperfectly estimated. Hence, improved transfer requires balancing the removal of pretrained priors against the preservation of discriminative forensic evidence. We next characterize this trade-off at the target-risk level using a local quadratic approximation.

Assumption A.2 (Local Target-Risk Surrogate). *Let $\mathcal{R}_{\text{tgt}}(\lambda)$ denote the target-domain risk under the suppressed representation $T_\lambda(\boldsymbol{f})$. Let $\boldsymbol{P}_{\mathcal{P}^\star}$ denote the ideal projection onto the pretraining-dominant subspace, and let $\boldsymbol{P}_{\hat{\mathcal{P}}}$ denote its estimated counterpart. In the operating regime of interest, we assume that the target-domain risk admits the following second-order surrogate:*

$$\mathcal{R}_{\text{tgt}}(\lambda) \approx \mathcal{R}_0 + (1 - \lambda)^2 \Gamma_{\text{prior}} + \lambda^2 \Gamma_{\text{share}} + \lambda^2 \Gamma_{\text{mis}}, \tag{10}$$

where

$$\Gamma_{\text{prior}} \triangleq \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{tgt}}} \left[\omega_{\text{prior}}(\boldsymbol{x}, y) \left\| \boldsymbol{P}_{\mathcal{P}^\star} \boldsymbol{f}_{\text{sem}}(\boldsymbol{x}) \right\|_2^2\right] > 0, \tag{11}$$

$$\Gamma_{\text{share}} \triangleq \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{tgt}}} \left[\omega_{\text{share}}(\boldsymbol{x}, y) \left\| \boldsymbol{P}_{\hat{\mathcal{P}}} \boldsymbol{f}_{\text{shr}}(\boldsymbol{x}) \right\|_2^2\right] > 0, \tag{12}$$

$$\Gamma_{\text{mis}} \triangleq \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{tgt}}} \left[\omega_{\text{mis}}(\boldsymbol{x}, y) \left\| \left(\boldsymbol{P}_{\hat{\mathcal{P}}} - \boldsymbol{P}_{\mathcal{P}^\star}\right) \boldsymbol{f}(\boldsymbol{x}) \right\|_2^2\right] \geq 0. \tag{13}$$

Here, Γ_{prior} represents the effective penalty of retaining pretrained priors, Γ_{share} the cost of suppressing shared task-relevant evidence, and Γ_{mis} the cost induced by subspace mismatch. The weights $\omega_{\text{prior}}, \omega_{\text{share}}$, and ω_{mis} absorb local loss curvature and sample-dependent sensitivity. Equation 14 is a local second-order surrogate motivated by equation 9. $\mathcal{R}_0$ denotes a residual base risk and should not be interpreted as $\mathcal{R}_{\text{tgt}}(0)$. Its role is to isolate the leading-order benefit of attenuating unstable pretrained priors from the cost of suppressing shared or misidentified task-relevant information.

Assumption A.3 (Local Target-Risk Surrogate)**.** *Let $\mathcal{R}_{\text{tgt}}(\lambda)$ denote the target-domain risk under the suppressed representation $T_\lambda(\boldsymbol{f})$. Let $\boldsymbol{P}_{\mathcal{P}^\star}$ denote the ideal projection onto the pretraining-dominant subspace, and let $\boldsymbol{P}_{\hat{\mathcal{P}}}$ denote its estimated counterpart. In the operating regime of interest, we assume that the target-domain risk admits the following second-order surrogate:*

$$\mathcal{R}_{\text{tgt}}(\lambda) \approx \mathcal{R}_0 + (1-\lambda)^2\Gamma_{\text{sem}} + \lambda^2\Gamma_{\text{share}} + \lambda^2\Gamma_{\text{mis}}, \tag{14}$$

where

$$\Gamma_{\text{sem}} \triangleq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{tgt}}}\left[\omega_{\text{sem}}(\boldsymbol{x},y)\,\|\boldsymbol{P}_{\mathcal{P}^\star}\boldsymbol{f}_{\text{sem}}(\boldsymbol{x})\|_2^2\right] > 0, \tag{15}$$

$$\Gamma_{\text{share}} \triangleq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{tgt}}}\left[\omega_{\text{share}}(\boldsymbol{x},y)\,\left\|\boldsymbol{P}_{\hat{\mathcal{P}}}\boldsymbol{f}_{\text{shr}}(\boldsymbol{x})\right\|_2^2\right] > 0, \tag{16}$$

$$\Gamma_{\text{mis}} \triangleq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{tgt}}}\left[\omega_{\text{mis}}(\boldsymbol{x},y)\,\left\|\left(\boldsymbol{P}_{\hat{\mathcal{P}}} - \boldsymbol{P}_{\mathcal{P}^\star}\right)\boldsymbol{f}(\boldsymbol{x})\right\|_2^2\right] \geq 0. \tag{17}$$

Here, Γ_{sem} represents the effective penalty of retaining pretrained semantic priors, Γ_{share} the cost of suppressing shared task-relevant evidence, and Γ_{mis} the cost induced by subspace mismatch. The weights ω_{sem}, ω_{share}, and ω_{mis} absorb local loss curvature and sample-dependent sensitivity. Equation 14 is a local second-order surrogate motivated by equation 9. $\mathcal{R}_0$ denotes a residual base risk and should not be interpreted as $\mathcal{R}_{\text{tgt}}(0)$. Its role is to isolate the leading-order benefit of attenuating unstable pretrained priors from the cost of suppressing shared or misidentified task-relevant information.

Proposition A.2 (Optimal Suppression under the Local Surrogate)**.** *Under Assumption A.3, the approximate target risk in equation 14 is uniquely minimized at*

$$\lambda^\star = \frac{\Gamma_{\text{sem}}}{\Gamma_{\text{sem}} + \Gamma_{\text{share}} + \Gamma_{\text{mis}}}. \tag{18}$$

Moreover, $\lambda^\star \in (0,1)$ whenever $\Gamma_{\text{sem}} > 0$ and $\Gamma_{\text{share}} + \Gamma_{\text{mis}} > 0$.

Proof. Let $\Gamma_{\text{cost}} = \Gamma_{\text{share}} + \Gamma_{\text{mis}}$. Ignoring the constant $\mathcal{R}_0$, equation 14 becomes

$$\widetilde{\mathcal{R}}_{\text{tgt}}(\lambda) = (1-\lambda)^2\Gamma_{\text{sem}} + \lambda^2\Gamma_{\text{cost}}. \tag{19}$$

Taking the derivative with respect to λ and setting it to zero gives

$$-2(1-\lambda)\Gamma_{\text{sem}} + 2\lambda\Gamma_{\text{cost}} = 0, \tag{20}$$

which yields equation 18. Since the second derivative is $2(\Gamma_{\text{sem}} + \Gamma_{\text{cost}}) > 0$, the minimizer is unique. The condition $\lambda^\star \in (0,1)$ follows directly from $\Gamma_{\text{sem}} > 0$ and $\Gamma_{\text{cost}} > 0$. □

Proposition A.2 characterizes the trade-off induced by pretrained-prior suppression under Assumption A.3. It provides an analytical interpretation of the empirical observation that stronger suppression can reduce the influence of pretrained semantic priors, while excessive suppression may also remove shared task-relevant information. The balance between these competing effects leads to an intermediate preferred suppression strength under the local target-risk approximation.

B PERTURBATION ANALYSIS OF INFERENCE COMPATIBILITY

We analyze the compatibility of different inference protocols with Batch-SVD training from the perspective of pretrained-prior subspace perturbation.

Assumption B.1 (Shared Pretrained-Prior Geometry)**.** *Let $P^\star$ denote the latent projector onto the pretraining-dominant subspace induced by the frozen CLIP anchor. We assume that the projectors estimated by three inference protocols, denoted by P_{batch}, P_{single}, and P_{ref}, respectively, are all perturbations of the same latent pretrained-prior geometry. Specifically, they satisfy*

$$\|P_\alpha - P^\star\| \le \delta_\alpha, \quad \alpha \in \{batch, single, ref\}, \tag{21}$$

for some bounded perturbation levels δ_α.

Proposition B.1 (Inference Gap under Projector Perturbation)**.** *Let*

$$T_\lambda^{(\alpha)}(\boldsymbol{f}) = (I - \lambda P_\alpha)\boldsymbol{f} \tag{22}$$

denote the decoupled feature produced by inference protocol $\alpha \in \{batch, single, ref\}$. Under Assumption B.1, for any two protocols α and β, their estimated projectors satisfy

$$\|P_\alpha - P_\beta\| \le \delta_\alpha + \delta_\beta. \tag{23}$$

Consequently, the resulting feature discrepancy is bounded by

$$\left\| T_\lambda^{(\alpha)}(\boldsymbol{f}) - T_\lambda^{(\beta)}(\boldsymbol{f}) \right\| \le \lambda(\delta_\alpha + \delta_\beta)\|\boldsymbol{f}\|. \tag{24}$$

Moreover, if the downstream score function h is L-Lipschitz, then

$$\left| h(T_\lambda^{(\alpha)}(\boldsymbol{f})) - h(T_\lambda^{(\beta)}(\boldsymbol{f})) \right| \le L\lambda(\delta_\alpha + \delta_\beta)\|\boldsymbol{f}\|. \tag{25}$$

Proof. By Assumption B.1 and the triangle inequality,

$$\|P_\alpha - P_\beta\| \le \|P_\alpha - P^\star\| + \|P_\beta - P^\star\| \le \delta_\alpha + \delta_\beta. \tag{26}$$

Moreover, by the definition of $T_\lambda^{(\alpha)}$,

$$\begin{aligned} T_\lambda^{(\alpha)}(\boldsymbol{f}) - T_\lambda^{(\beta)}(\boldsymbol{f}) &= (I - \lambda P_\alpha)\boldsymbol{f} - (I - \lambda P_\beta)\boldsymbol{f} \\ &= \lambda(P_\beta - P_\alpha)\boldsymbol{f}. \end{aligned} \tag{27}$$

Taking norms gives

$$\left\| T_\lambda^{(\alpha)}(\boldsymbol{f}) - T_\lambda^{(\beta)}(\boldsymbol{f}) \right\| \le \lambda\|P_\alpha - P_\beta\|\|\boldsymbol{f}\| \le \lambda(\delta_\alpha + \delta_\beta)\|\boldsymbol{f}\|. \tag{28}$$

The prediction-score bound follows directly from the L-Lipschitz continuity of h. □

Proposition B.1 shows that the discrepancy across inference protocols is bounded by the variation in their estimated pretrained-prior projectors. From this perspective, Batch-SVD can be viewed as a training-time approximation to pretrained-prior subspace estimation. Batch Inference, Per-sample Inference, and Reference-based Inference mainly differ in how the pretrained-prior subspace is estimated, while sharing the same anchor-induced representation geometry. Therefore, when their estimated projectors remain sufficiently close, the resulting decoupled features are also close to those produced by Batch Inference.

C More Experiments

C.1 Robustness Evaluation

To assess the resilience of our model against real-world image distortions, we follow the rigorous evaluation protocols established in prior studies (Haliassos et al., 2021; Yan et al., 2024b; Jiang et al., 2020). The assessment covers five distinct degradation categories: Color Saturation, JPEG Compression, Color Contrast, Block-wise noise, and Gaussian Blur. For each category, we apply distortions ranging from severity level 0 to 5 to measure performance stability. All models are trained on the FaceForensics++ (c23) dataset and evaluated using video-level AUC. We benchmark our approach against leading methods such as Face X-ray (Li et al., 2020a), SBI (Shiohara & Yamasaki, 2022), LipForensics (Haliassos et al., 2021), and FWA (Li & Lyu, 2019). To ensure a controlled comparison, we also reproduced F3Net (Qian et al., 2020), UCF (Yan et al., 2023a), and CORE (Ni et al., 2022) using **the same data augmentation pipelin**e employed in our framework.

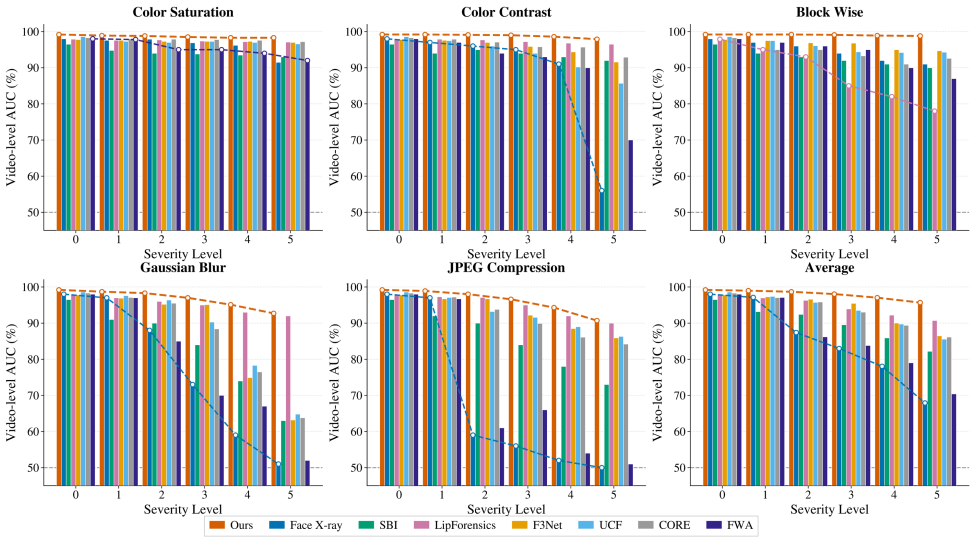


Figure 5: Robustness analysis against common image corruptions. The curves illustrate the Video-level AUC (%) performance of our method versus existing approaches. We test across five different degradation scenarios, including Color Saturation, JPEG Compression, Color Contrast, Block-wise noise, and Gaussian Blur, plotted against increasing severity levels (0-5).

The quantitative results are presented in Figure 5. As shown in the "Average" plot, our method consistently secures the top performance across nearly all severity levels, outperforming both the standard baselines and the reproduced counterparts sharing identical augmentation schemes. While competitors experience sharp performance declines under severe distortions, specifically in JPEG compression and Gaussian blur where high-frequency details are compromised, our model exhibits superior stability. This evidence confirms that the robustness of our method is intrinsic to the learned generalized features rather than being a byproduct of data augmentation.

Table 11: Effect of the dimension k on the testing datasets. Frame-level AUC (%) is reported.

k	CDv2	DFD	DFDC	SimSwap	UniFace	Avg.
1	89.8	94.0	**85.9**	**91.4**	95.9	91.4
2	90.2	94.0	85.5	91.3	95.9	91.4
8	90.5	94.4	**85.9**	91.2	96.3	**91.7**
16	90.4	94.4	85.7	91.0	**96.4**	91.6
32	90.5	94.4	85.6	91.0	**96.4**	91.6
64	**90.6**	**94.5**	85.8	91.1	96.3	**91.7**

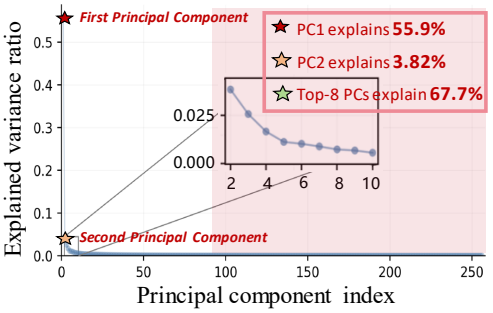


Figure 6: Explained variance ratio of principal components. The pretrained-prior variance is concentrated in the first few principal components, particularly the first and second ones.

C.2 The Impact of the Estimated Subspace Dimension k.

We further investigate the effect of the pretraining-dominant subspace dimension k, as reported in Table 11. Even a very small subspace already yields strong performance: $k = 1$ remains competitive across both cross-domain and cross-manipulation benchmarks, indicating that removing the most dominant pretrained-prior direction is already beneficial. Performance further improves as k increases to 8, while larger values bring only marginal gains. Nevertheless, to ensure robustness across diverse cross-domain settings, we use $k = 64$ as the default.

We further analyze why small values of k, *e.g.*, $k = 1$ or $k = 2$, are already effective. As shown in Figure 6, using 5,000 randomly sampled FF++ images, the pretrained-prior variance captured by the frozen CLIP features is highly concentrated in the first few principal components, particularly the first two, while later components contribute relatively little. This low-rank structure explains the

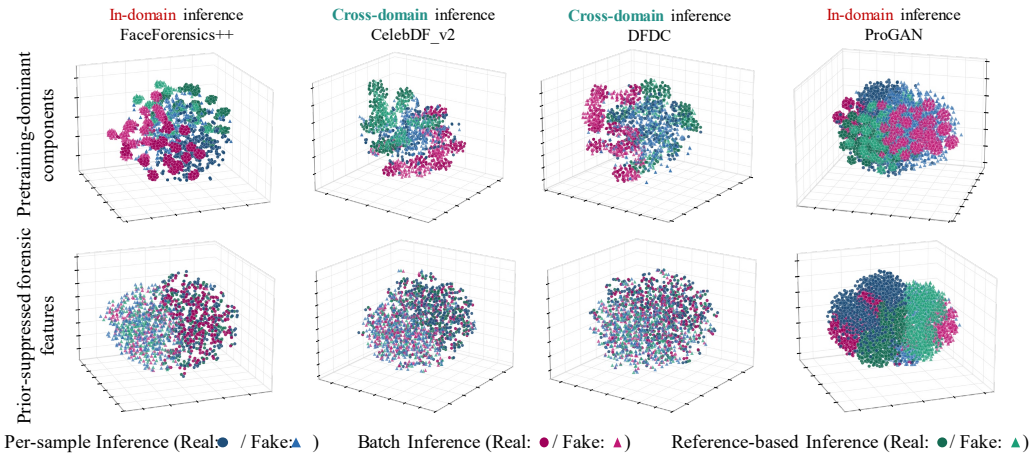


Figure 7: **Feature visualization of different GSD inference strategies.** We visualize 3D t-SNE feature distributions under Per-sample Inference, Batch Inference, and Reference-based Inference. The top row shows the pretraining-dominant components estimated from the frozen CLIP anchor, while the bottom row shows the resulting forensic features after pretrained-prior suppression. Across the three inference strategies, the pretraining-dominant components occupy highly similar feature spaces, while the prior-suppressed forensic features exhibit consistent reorganization into comparable forensic-oriented feature spaces.

trend in Table 11: once the dominant pretrained-prior directions are suppressed, further increasing k provides diminishing returns.

C.3 Analysis of initial settings for parameters a and b

We further analyze the effect of initializing the adaptive modulation parameters a and b, which determine the mapping from pretrained geometry occupancy r to suppression strength λ. As shown in Table 12 and Figure 8, $(2, 1)$ achieves the best overall performance, while the relatively small differences across settings indicate that the adaptive mechanism is robust to initialization. The learned modulation also exhibits a compensatory behavior: stronger initial suppression is moderated during training, whereas weaker initialization is adjusted upward. This trend is consistent with the target-risk trade-off discussed in Section 2.4, where intermediate suppression balances pretrained-prior interference against the preservation of discriminative forensic evidence. Moreover, the learned mappings vary across layers, with deeper layers generally exhibiting stronger suppression. This observation is consistent with the stronger pretraining-dominant structure observed in deeper CLIP representations and supports the use of layer-specific adaptive coefficients.

Table 12: Effect of different initialization settings for a and b. ACC (%) is reported on the test datasets.

(a, b)	CDv2	DFDC	DFDCP	SimSwap	FSGAN	Avg.
$(1, 0)$	81.9	76.9	81.1	**84.2**	**90.9**	83.0
$(1, 1)$	82.8	76.6	81.9	84.0	90.8	83.2
$(2, 1)$	**83.1**	**77.0**	**82.1**	83.9	90.8	**83.4**
$(2, -1)$	**83.1**	76.8	**82.1**	83.9	90.6	83.3
$(6, -3)$	82.3	**77.0**	81.4	83.8	90.7	83.0

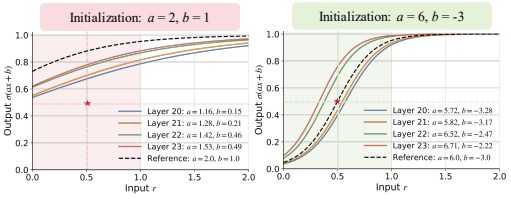


Figure 8: Adaptive suppression curves under different initializations of a and b.

C.4 The Impact of Different Batch Sizes

Considering practical deployment constraints, we evaluate the effects of both training and inference batch sizes, as reported in Table 13. Reducing the training batch size from 128 to 16 only slightly

decreases the average frame-level AUC from 92.3% to 92.0%, indicating that larger batches provide more stable semantic-subspace estimation but are not essential for strong performance. The effect of inference batch size is even smaller: with the default training batch size, varying it from 32 to 256 produces almost no performance change. Overall, GSD is robust to batch-size variation, enabling flexible deployment under limited GPU memory and computational resources.

C.5 Feature Visualization of Different Reference Strategies

We examine whether different inference protocols preserve the pretrained-prior suppression effect of GSD through 3D t-SNE visualizations in Figure 7. We consider both face-forgery and natural-image synthesis detection, covering in-domain settings including FaceForensics++ and ProGAN, and cross-domain settings including CelebDF_v2 and DFDC. As shown in the top row, the pretrained-prior components estimated by the three inference protocols occupy highly similar feature spaces. Despite differences in data domain and visual content, they form comparable structured manifolds, indicating that Per-sample, Batch, and Reference-based Inference recover similar pretraining-dominant directions from the frozen CLIP anchor. This suggests that the pretrained-prior subspace estimated by GSD remains stable across inference protocols, consistent with our analysis in Section 3.3. More importantly, the bottom row shows that, after pretrained-prior suppression, the resulting forensic features exhibit highly similar reorganization patterns across the three protocols. Compared with the original pretrained-prior components, these features are less organized by identity, object category, or scene-level semantics, while occupying comparable forensic-oriented feature spaces. This indicates that Reference-based Inference preserves the essential pretrained-prior suppression geometry of GSD rather than inducing a substantially different representation space from Per-sample and Batch Inference.

Table 13: Effect of training and inference batch sizes on frame-level AUC (%).

Batchsize	CDv1	DFD	DFDC	FaceDancer	FSGAN	Avg.
Different Training Batch Sizes (Test batch size $4\times$)						
$B_{tr} = 16$	90.6	94.4	85.0	93.6	96.3	92.0
$B_{tr} = 32$	90.5	94.2	85.1	93.7	96.4	92.0
$B_{tr} = 64$	90.9	94.1	86.1	93.6	96.4	92.2
$B_{tr} = 128$	91.3	94.5	85.8	93.8	96.1	92.3
Different Inference Batch Sizes (Training Batch Size = 128)						
$B_{ts} = 32$	91.1	94.4	85.8	93.8	96.1	92.2
$B_{ts} = 64$	91.2	94.5	85.8	93.9	96.1	92.3
$B_{ts} = 128$	91.0	94.5	85.5	93.9	96.1	92.2
$B_{ts} = 256$	91.3	94.5	85.7	93.8	96.1	92.3

C.6 More Attention VISUALIZATION

We further examine whether GSD encourages the detector to capture forgery-related evidence by visualizing attention maps on four datasets with localized facial manipulations, including FaceForensics++, Celeb-DF-v2, DFDC, and FaceDancer. As shown in Fig. 9, the naively fine-tuned CLIP exhibits highly similar attention patterns for real and fake images across different datasets, with only limited variation across different manipulated regions. In contrast, GSD-CLIP shows a clear distinction between real and fake images after suppressing the semantic priors. *For fake images, its attention is closely aligned with the manipulated areas, despite being trained without any region-level supervision or manipulation masks. For real images, where such forgery evidence is absent, the attention is instead more diffuse across the image.* These observations provide qualitative evidence that GSD enables the detector to discover forensic cues with reduced interference from pretrained semantic priors.

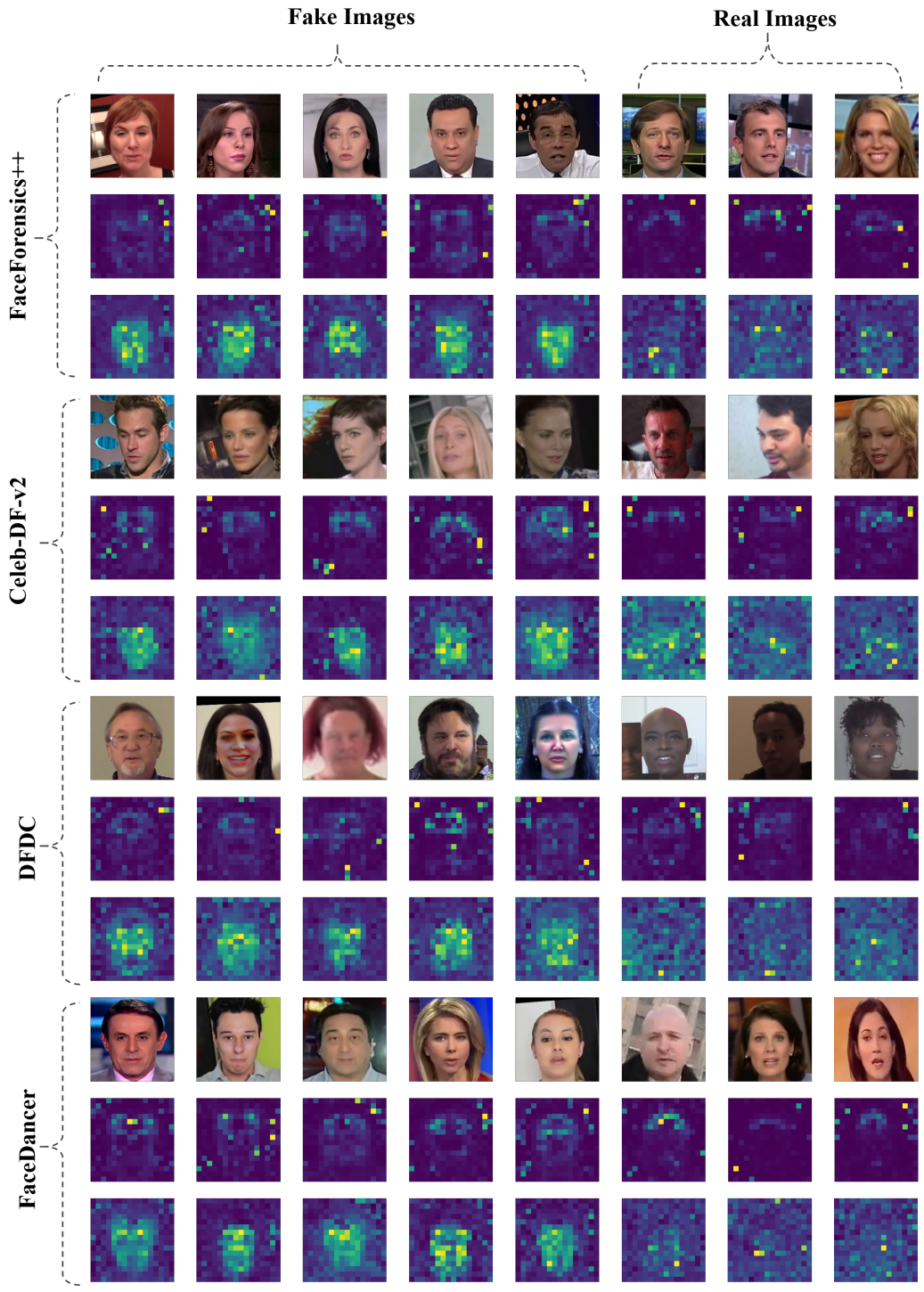


Figure 9: Attention visualization of the naively fine-tuned CLIP and GSD-CLIP on four datasets with localized facial manipulations. From top to bottom within each group: input images, naively finetuned CLIP, and GSD-CLIP. While naive fine-tuning yields similar face-dominant attention for all images, GSD produces more localized responses around potential manipulation regions for fake images and more diffuse responses for real images.