CIFT: DONOR-GROUNDED RELATIONAL GAP LEARNING FOR SOURCE-FREE FACE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Face-forgery detectors often generalize poorly across synthesis pipelines because visual artifacts change with the renderer. We study whether source information available during training can provide a more transferable supervisory signal without being required at deployment. We introduce CIFT (**C**onsistency-**I**nvariant **F**orensic **T**raining), a privileged-learning framework that uses a donor reference only during training. For swaps, the reference is the identity donor; for reenactment, it is the driving/source reference and defines a weaker relational discrepancy rather than identity replacement. XID-Mamba separately encodes the observed and donor streams with bidirectional Mamba modeling and symmetric cross-attention, while Type-Conditioned IGS structures their relation through complementary directional and magnitude constraints. Auxiliary diffusion regularizes the stream representations. At inference, all donor-dependent relational and diffusion components are discarded, leaving a single-image source-free detector. Trained only on FaceForensics++ (c23), CIFT generalizes strongly across five held-out benchmarks, remains effective under an encoder- and diffusion-prior-aligned comparison with DiffusionFake, and degrades under donor-correspondence interventions that replace the true donor with non-corresponding alternatives.

1 INTRODUCTION

Face-forgery generation has progressed from GAN-based face swapping and neural-rendering-based reenactment to latent-diffusion synthesis Rössler et al. (2018); Thies et al. (2019); Rombach et al. (2022); Zhao et al. (2023), enabling increasingly realistic identity impersonation, fraud, nonconsensual imagery, and disinformation Rössler et al. (2018); Li et al. (2020); Zi et al. (2024). The central challenge is therefore not only detecting known manipulations, but generalizing when the synthesis pipeline changes. Existing detectors often rely on spatial, frequency, blending, or reconstruction artifacts Qian et al. (2020); Cao et al. (2022); Shiohara & Yamasaki (2022); Sun et al. (2021b), which can shift across manipulation methods. Identity cues can also induce training-specific correlations Dong et al. (2023), motivating a distinction between useful relational identity supervision and direct subject memorization. We therefore study a deliberately scoped question: whether training-only donor supervision improves single-image cross-pipeline generalization, complementary to temporal and audio-visual modeling.

Face swapping provides a structural relation during training: donor-conditioned facial appearance is rendered within target-conditioned context. CIFT does not assume that a source-free test image exposes two separately recoverable identities, nor does inference attempt to reconstruct the donor. Instead, the donor reference acts as privileged supervision for the same target representation used by the deployed classifier. During training, relational losses computed from the observed and donor streams back-propagate through the target stream and shared encoder, thereby shaping f_t while the classification objective is applied directly to f_t. At inference, the donor branch is removed and the detector operates from x alone without evaluating the training-time relation.

Throughout the paper, x_s denotes the training-time donor reference. For swaps, Δ represents the donor–target identity gap; for reenactment, it represents the weaker driving–target discrepancy above. Existing identity-aware methods instead operate on the observed face: ICT Dong et al. (2022) and SELFI Kim et al. (2025) assess internal identity consistency, CADDM Dong et al. (2023) suppresses identity leakage, and IID Huang et al. (2023) models explicit–implicit identity discrepancy.

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.

CIFT differs by using the actual donor reference as a training-only relational input whose geometry supervises the representation retained by the source-free detector.

We introduce CIFT (**C**onsistency-**I**nvariant **F**orensic **T**raining), a privileged-learning framework Vapnik & Izmailov (2015) that uses source correspondence only during training while preserving source-free inference. The full training configuration assumes a reference x_s and, for forged samples, a coarse relation type τ; neither is required at deployment. XID-Mamba processes donor and target streams with bidirectional Mamba Gu & Dao (2024) scans and symmetric cross-attention to produce embeddings g_s and g_t and the donor-grounded relational gap $\Delta{=}\|g_s - g_t\|_2$, whose identity-gap interpretation applies to face swaps. Type-Conditioned IGS structures this relation through complementary directional and magnitude constraints. A Stable Diffusion-based pathway provides auxiliary training regularization. At inference, the donor stream, diffusion pathway, and XID-Mamba/IGS machinery are removed; only the Global Head and Spatial Mamba Classifier (SMC) process the observed face. Motivated by reducing dependence on generator-specific variation, consistent with domain-adaptation intuition Ben-David et al. (2010), we test CIFT through cross-dataset evaluation and controlled interventions. Encoder-aligned comparisons test whether the gain persists when the encoder family is held fixed, component ablations separate donor-side diffusion supervision from explicit XID-Mamba modeling and IGS, and correspondence controls test whether the true donor assignment itself carries useful supervision.

Contributions. Our contributions are threefold:

- We introduce **donor-grounded privileged learning** for source-free face-forgery detection, using a donor reference only during training while requiring only the observed face at inference.
- We propose **XID-Mamba with Type-Conditioned IGS** to construct and structure a donor-grounded relational gap through complementary directional and magnitude constraints, with the identity-gap interpretation applying to face swaps.
- We validate donor-grounded supervision through cross-dataset evaluation, backbone and component controls, and donor-correspondence controls that test dependence on the structured donor–target relation.

2 Related Work

Generalizable face-forgery detection. Existing detectors improve cross-domain generalization through frequency analysis, reconstruction, feature disentanglement, attention, blending simulation, and domain reweighting Qian et al. (2020); Cao et al. (2022); Sun et al. (2021b); Yan et al. (2023); Zhao et al. (2021); Shiohara & Yamasaki (2022); Sun et al. (2021a). These approaches primarily learn visual evidence from the observed forgery, whose appearance can vary with the synthesis pipeline. CIFT instead investigates whether donor–target relations available during training can provide complementary supervision for a detector that remains source-free at deployment.

Identity-aware detection. Identity provides a complementary signal for generalization. CADDM Dong et al. (2023) suppresses implicit identity leakage, ICT Dong et al. (2022) measures consistency between facial regions, and SELFI Kim et al. (2025) uses a pretrained face recognizer to assess internal identity consistency. Most closely related, IID Huang et al. (2023) detects face swaps through the discrepancy between explicit identity, obtained from face recognition, and an implicit identity representation learned from the observed face. Its contribution establishes identity discrepancy as a useful forensic signal.

Table 1: Comparison with identity-aware and diffusion-guided detectors. "Donor relational input" denotes a separately encoded donor reference. "Donor–target gap" denotes an explicit distance constructed between separately encoded donor-reference and observed-face embeddings.

Method	Donor relational input x_s	Donor–target gap Δ	Source-free inference
ICT Dong et al. (2022)	✗	✗	✓
CADDM Dong et al. (2023)	✗	✗	✓
IID Huang et al. (2023)	✗	✗	✓
SELFI Kim et al. (2025)	✗	✗	✓
DiffusionFake Sun et al. (2024)	✗	✗	✓
CIFT (Ours)	✓	✓	✓

CIFT does not claim identity discrepancy itself as new. Unlike IID Huang et al. (2023), which derives explicit and implicit identity representations from the observed face, CIFT separately encodes the actual donor reference and observed face during training and directly supervises their relational

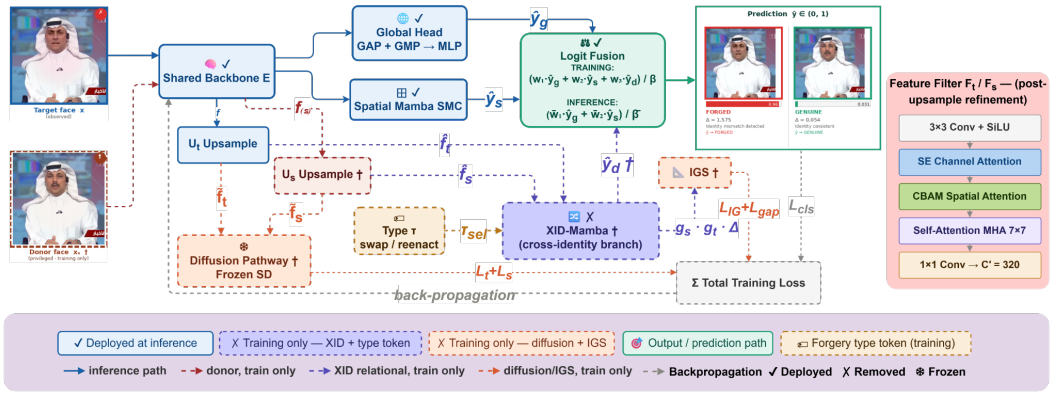

Figure 1: Overview of CIFT. During training, the observed face x and its donor reference x_s share encoder E. The deployed Global Head and SMC classify x, while the training-only XID-Mamba branch constructs an explicit donor–target relation that is structured by IGS. A Stable Diffusion pathway provides auxiliary representation regularization. At inference, the donor stream, XID-Mamba, IGS, and diffusion pathway are removed.

geometry through XID-Mamba and IGS. The donor reference is removed at deployment. The novelty is therefore **donor-grounded privileged relational supervision**, not identity-aware detection itself.

Diffusion-guided forensics. DiffusionFake Sun et al. (2024) uses a frozen Stable Diffusion model Rombach et al. (2022) and source and target images to guide reconstruction. Its source- and target-related feature streams are, however, both transformed from the encoded observed forgery. CIFT retains diffusion guidance as auxiliary representation regularization but separately encodes the actual donor reference and observed face to construct an explicit relational representation. The distinction is therefore not whether source information is available during training, but whether the actual donor reference is separately encoded as a relational input whose geometry is directly supervised.

Selective state-space models. Mamba Gu & Dao (2024) provides input-dependent selective state-space modeling with linear sequence complexity and has been adapted to visual representations Hatamizadeh & Kautz (2025); Zhu et al. (2024). CIFT uses Mamba for spatial modeling in SMC and for donor–target relational modeling in XID-Mamba. The privileged supervision itself arises from the donor–target relation, rather than from the state-space architecture.

Privileged information learning. Learning using privileged information provides auxiliary signals during training that are unavailable at test time Vapnik & Izmailov (2015). CIFT follows this setting by using the donor reference as a relational input only during training and removing all donor-dependent computation at deployment. Knowledge distillation Hinton et al. (2015) provides an alternative route for transferring privileged information through teacher outputs, whereas CIFT directly supervises the geometry of the donor–target relation through XID-Mamba and IGS.

3 Method

3.1 Overview and Problem Formulation

Let $x \in \mathbb{R}^{3\times H\times W}$ denote the observed face and $y \in \{0, 1\}$ its label, where 0 denotes genuine and 1 denotes forged. During training, each manipulated sample is associated with a donor reference x_s; at inference, only x is available. Throughout this paper, **source-free inference** means that no donor image, donor identity, or correspondence metadata is required at deployment.

For face swapping, x_s is the identity donor associated with $\mathcal{I}_s$, while x retains target-conditioned context associated with $\mathcal{I}_t$. For reenactment, x_s denotes the driving/source reference; under our donor-reference convention, this does not imply identity replacement and instead defines a weaker relational discrepancy. For genuine samples, x_s is a temporally distinct valid frame of the same

identity as x. Under the default training construction, each relational pair therefore contains two distinct images; Table 7 separately defines controlled donor-assignment interventions such as $x_s = x$.

The observed and donor-reference faces share an encoder E and projection ϕ:

$$f_t = \phi(E(x)), \qquad f_s = \phi(E(x_s)), \qquad \hat{f}_t = U_t(f_t), \qquad \hat{f}_s = U_s(f_s), \tag{1}$$

where U_t and U_s are independent stream transformations. The source-free detector operates directly on f_t through a Global Head and a Spatial Mamba Classifier (SMC). The transformed pair $(\hat{f}_t, \hat{f}_s)$ is used only during training to construct the donor–target relation described below.

3.2 Donor-Grounded Relational Modeling

The central operation in CIFT is XID-Mamba, which explicitly relates the donor reference and observed target rather than inferring the relation solely from the observed image. Each stream is pooled to a 7×7 token grid, projected to dimension $d = 512$, and processed by an independent bidirectional Mamba encoder Gu & Dao (2024). For forged samples, a learned type token indicating swap or reenactment is prepended before the scan; genuine pairs use no manipulation token. Writing $\mathcal{R}_\tau$ for this routing operation,

$$\begin{aligned} \mathbf{Z}_t &= \mathrm{BiMamba}_t\Big(\mathcal{R}_\tau\left(P_t(\mathrm{pool}_{7\times7}(\hat{f}_t))\right)\Big), \\ \mathbf{Z}_s &= \mathrm{BiMamba}_s\Big(\mathcal{R}_\tau\left(P_s(\mathrm{pool}_{7\times7}(\hat{f}_s))\right)\Big). \end{aligned} \tag{2}$$

The two contextualized streams then interact through symmetric cross-attention:

$$\begin{aligned} \tilde{\mathbf{Z}}_t &= \mathrm{LN}(\mathbf{Z}_t + \mathrm{MHA}(\mathbf{Z}_t, \mathbf{Z}_s, \mathbf{Z}_s)), \\ \tilde{\mathbf{Z}}_s &= \mathrm{LN}(\mathbf{Z}_s + \mathrm{MHA}(\mathbf{Z}_s, \mathbf{Z}_t, \mathbf{Z}_t)). \end{aligned} \tag{3}$$

Mean pooling produces donor and target embeddings $g_s, g_t \in \mathbb{R}^{512}$. Their Euclidean distance defines the type-conditioned donor-grounded relational gap:

$$g_t = \mathrm{mean}(\tilde{\mathbf{Z}}_t), \qquad g_s = \mathrm{mean}(\tilde{\mathbf{Z}}_s), \qquad \Delta = \|g_s - g_t\|_2. \tag{4}$$

For swaps, Δ represents the donor–target identity gap; for reenactment, it represents the weaker driving–target discrepancy above. Table 5 evaluates donor-grounded XID-Mamba training against the source-free detector, while Table 7 separately tests whether its benefit depends on the correct donor–target correspondence.

3.3 Type-Conditioned Identity Gap Supervision

IGS structures the donor–target relation along complementary directional and magnitude dimensions. For swaps, this relation has an identity-gap interpretation; for reenactment, it represents the weaker driving–target discrepancy defined above. Let $d_{\cos} = 1 - \bar{g}_t^\top \bar{g}_s$, where $\bar{g}_t$ and $\bar{g}_s$ are normalized embeddings. The directional objective is

$$\mathcal{L}_{\mathrm{IG}} = (1-y)\big[\max(0, d_{\cos} - \delta)\big]^2 + y\big[\max(0, m(\tau) - d_{\cos})\big]^2, \tag{5}$$

with genuine pull tolerance $\delta = 0.3$ and type-conditioned margins $m(\tau) = 1.0$ for swap and 0.8 for reenactment.

The complementary Euclidean objective supervises the magnitude of Δ:

$$\mathcal{L}_{\mathrm{gap}} = (1-y)\Delta^2 + y\big[\max(0, m'(\tau) - \Delta)\big]^2, \tag{6}$$

where $m'(\tau) = 1.2$ for swap and 0.9 for reenactment. Thus, $\mathcal{L}_{\mathrm{IG}}$ constrains relative direction, while $\mathcal{L}_{\mathrm{gap}}$ constrains separation magnitude. Table 6 evaluates the two objectives separately, jointly, and with type-conditioned margins.

3.4 Training and Source-Free Inference

The deployed detector combines complementary global and spatial evidence. The Global Head applies average and max pooling to f_t, while SMC uses bidirectional selective scanning over its spatial token grid. Their logits $\hat{y}_g$ and $\hat{y}_s$ form the source-free prediction

$$\hat{y}_{\mathrm{sf}} = \frac{\bar{w}_1 \hat{y}_g + \bar{w}_2 \hat{y}_s}{\bar{\beta}}. \tag{7}$$

Detailed branch definitions are given in Appendix A.3.

During training, binary classification is always performed by the same Global Head and SMC retained at inference. XID-Mamba and IGS do not produce classification logits. Instead, their relational losses back-propagate through the target stream and shared encoder, so the privileged donor signal shapes the target representation f_t on which the source-free classifier is trained. Consequently, deployment does not require recovering the donor identity or reproducing the relational gap: only the representation learned under this auxiliary supervision is retained. Manipulation type τ conditions only the training-time relational representation and IGS margins and is never an input to a classification head. A Stable Diffusion 1.5 Rombach et al. (2022) pathway provides auxiliary per-stream denoising regularization. All auxiliary components are removed at inference. The complete objective is

$$\mathcal{L} = \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{diff}}\mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{ig}}\mathcal{L}_{\mathrm{IG}} + \lambda_{\mathrm{gap}}\mathcal{L}_{\mathrm{gap}} + \mathcal{L}_{\mathrm{reg}}, \tag{8}$$

where, for the full donor-enabled configuration, $\mathcal{L}_{\mathrm{diff}} = \mathcal{L}_t + \mathcal{L}_s$, $\mathcal{L}_{\mathrm{cls}}$ is specified in Appendix B.2, and $\mathcal{L}_{\mathrm{reg}}$ is the cross-attention regularizer detailed in Appendix A.

At inference, x_s, the donor stream, XID-Mamba, IGS, and diffusion pathway are discarded. Only E, ϕ, the Global Head, and SMC remain. The deployed model therefore uses x alone and neither reconstructs the source nor evaluates the training-time gap Δ; the privileged relation affects inference only through the learned target representation. The same interface applies across encoder families (Table 3); Algorithm 1 summarizes training and inference.

4 Experiments

4.1 Experimental Setup

Training protocol. We train CIFT and all controlled variants exclusively on FaceForensics++ (FF++) Rössler et al. (2018) at compression level c23. The training set contains genuine YouTube videos and four manipulation families: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. For each forged frame, FF++ provides the corresponding source video used to construct the training-only reference x_s: the identity donor for swaps and the driving/source reference for reenactment. We select the temporally nearest valid source frame. The FF++ manipulation family also supplies the coarse training-time relation label $\tau \in \{\texttt{swap}, \texttt{reenact}\}$ used only by XID-Mamba and IGS. For genuine samples, x_s is the nearest distinct valid frame from the same video and identity and no relation label is used. Neither x_s nor τ is required at inference.

Cross-dataset evaluation. We evaluate without fine-tuning on five primary held-out benchmarks: Celeb-DF Li et al. (2020), WildDeepfake Zi et al. (2024), DFDC-Preview Dolhansky et al. (2020), DFD Dufour & Gully (2019), and DiffSwap Zhao et al. (2023). These benchmarks cover controlled and in-the-wild acquisition conditions as well as GAN, reenactment, and diffusion-based synthesis. DiffSwap uses the $6{,}000$ image evaluation partition and is subject-disjoint from the FF++ training identities. At test time, every CIFT prediction is computed from the observed face alone using Eq. 7; no donor, temporal, or audio input is used. Our primary scope is source-free frame-level visual detection, evaluated with AUC (↑) and EER (↓). Appendix C.10 additionally evaluates frozen checkpoints zero-shot on KoDF Kwon et al. (2021), Celeb-DF++ Li et al. (2025), and DiffusionFace Chen et al. (2024) as a broader contemporary generalization check.

Implementation and deployment cost. The default model uses ConvNeXt-V2-Base Woo et al. (2023) with 256×256 input resolution. Full training uses AdamW Loshchilov & Hutter (2019), OneCycleLR, and bf16 distributed data parallelism for 80 epochs on $4\times$ NVIDIA A100 80 GB GPUs, requiring approximately 12 hours. This cost includes the training-only donor, XID-Mamba/IGS, and diffusion pathways. None is executed at deployment: inference retains only E, ϕ, the Global Head,

and SMC and requires neither donor retrieval nor diffusion computation. Full implementation details are provided in Appendices B.1 and B.2.

Baselines. We compare against ten representative detectors covering artifact learning, reconstruction, disentanglement, blending simulation, identity suppression, and diffusion guidance: Xception Chollet (2017), F3-Net Qian et al. (2020), MAT Zhao et al. (2021), LTW Sun et al. (2021a), DCL Sun et al. (2021b), UCF Yan et al. (2023), SBI Shiohara & Yamasaki (2022), RECCE Cao et al. (2022), CADDM Dong et al. (2023), and DiffusionFake Sun et al. (2024). IID Huang et al. (2023) and SELFI Kim et al. (2025) are included where directly comparable results are available. Results for prior detectors in Tables 2 and 10 follow the frame-level FF++ (c23) evaluation of Sun et al. (2024); other experiments use the protocol stated in the corresponding section.

IID Huang et al. (2023) is the closest identity-based method: it measures the discrepancy between explicit and implicit identity within the observed face, whereas CIFT encodes the donor reference as a separate, training-only relational input. DiffusionFake Sun et al. (2024) is the closest diffusion-guided method and pairs an EfficientNet-B4 encoder with a frozen Stable Diffusion 1.5 prior. To separate donor-grounded supervision from the choice of encoder, we also report **EfficientNet-B4 + Ours**, which adopts the same encoder family and frozen SD 1.5 prior while adding XID-Mamba/IGS supervision. The full ConvNeXt-V2-B configuration is reported as CIFT.

Controlled studies. The experiments are organized to isolate the proposed mechanism rather than only compare final models. We first evaluate cross-dataset generalization and transfer across four backbone families. We then fix the evaluation setting and progressively examine the deployed source-free architecture, donor-side diffusion supervision, explicit XID-Mamba modeling, joint IGS, and type-conditioned IGS margins. Finally, donor-correspondence controls alter only the training-time donor assignment while preserving the source-free inference graph. All controlled variants use one FF++-validation-selected optimization configuration without variant-specific tuning on the reported benchmarks. Unless stated otherwise, these ablations report Celeb-DF, DiffSwap, and FF++ AUC; their three-dataset average is distinct from the five-benchmark average in Table 2.

5 RESULTS

Table 2: Cross-dataset generalization under FF++ (c23) training (AUC↑ / EER↓). Reported results use source-free evaluation without fine-tuning. "EfficientNet-B4 + Ours" uses the DiffusionFake-aligned encoder and frozen SD 1.5 prior; "ConvNeXt-V2-B + Ours" is the full model. Avg. is computed over the five reported benchmarks. **Bold**: best; underline: second-best.

Method	Celeb-DF		WildDeepfake		DFDC-P		DFD		DiffSwap		Avg.	
	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
Xception Chollet (2017)	65.27	38.77	66.17	40.14	69.80	35.41	87.86	21.04	74.25	32.04	72.67	33.48
F3-Net Qian et al. (2020)	71.21	34.03	67.71	40.17	72.88	33.38	86.10	26.17	76.89	30.83	74.96	32.92
MAT Zhao et al. (2021)	70.65	35.83	70.15	36.53	67.34	38.31	87.58	21.73	79.93	27.77	75.13	32.03
LTW Sun et al. (2021a)	77.14	29.34	67.12	39.22	74.58	33.81	88.56	20.57	77.95	29.01	77.07	30.39
DCL Sun et al. (2021b)	82.30	26.53	71.14	36.17	76.71	31.97	91.66	16.63	80.21	27.37	80.40	27.73
SBI Shiohara & Yamasaki (2022)	80.76	26.97	68.22	38.11	76.53	30.22	88.13	17.25	75.20	31.49	77.77	28.81
RECCE Cao et al. (2022)	70.50	35.34	67.93	39.82	75.88	32.41	89.91	19.95	77.59	29.38	76.36	31.38
CADDM Dong et al. (2023)	77.56	30.63	72.56	33.63	72.45	33.56	82.90	25.20	75.58	31.01	76.21	30.81
UCF Yan et al. (2023)	82.44	25.31	73.21	34.89	75.63	31.74	90.12	18.43	79.87	27.95	80.25	27.66
IID Huang et al. (2023)	83.80	24.85	–	–	81.23	26.80	<u>93.92</u>	<u>14.01</u>	–	–	–	–
DiffusionFake Sun et al. (2024)	83.17	24.59	75.17	33.25	77.35	30.17	91.71	16.27	82.02	25.55	81.88	25.97
SELFI Kim et al. (2025)	83.90	–	–	–	84.00	–	90.70	–	–	–	–	–
EfficientNet-B4 + Ours	<u>88.40</u>	<u>19.83</u>	<u>84.10</u>	<u>23.79</u>	<u>84.17</u>	<u>23.05</u>	91.13	16.25	<u>84.03</u>	<u>23.85</u>	<u>86.37</u>	<u>21.35</u>
ConvNeXt-V2-B + Ours (CIFT)	**95.30**	**12.05**	**92.98**	**14.66**	**90.38**	**17.05**	**94.64**	**12.78**	**98.47**	**5.91**	**94.35**	**12.49**

Appendix C provides supporting analyses on cross-manipulation transfer, perturbation robustness, relational-head and complexity controls, auxiliary diffusion supervision, type-token injection, seed stability, and video-level evaluation.

5.1 CROSS-DATASET GENERALIZATION

Table 2 evaluates whether donor-grounded training transfers to source-free inference on unseen datasets. Trained only on FF++ (c23), CIFT achieves **94.35** average AUC and **12.49** average EER across five held-out benchmarks, compared with 81.88 AUC and 25.97 EER for DiffusionFake Sun

et al. (2024). This corresponds to a $+12.47$-point AUC difference and a 13.48-point EER reduction, with CIFT obtaining the highest AUC among the methods listed in Table 2. Under the encoder- and diffusion-prior-aligned EfficientNet-B4 comparison, CIFT reaches 86.37 average AUC, improving DiffusionFake by $+4.49$ points and outperforming it on four of five benchmarks. The gain therefore persists under this aligned comparison, while the full ConvNeXt-V2-B configuration further increases average AUC to 94.35, showing that backbone choice also contributes to final system performance. Among identity-based methods, IID Huang et al. (2023) reports $83.80/24.85$, $81.23/26.80$, and $93.92/14.01$ AUC/EER on Celeb-DF, DFDC-P, and DFD, respectively. CIFT reaches $95.30/12.05$, $90.38/17.05$, and $94.64/12.78$ on the same benchmarks. Because architectures and preprocessing differ, these figures serve as benchmark comparisons rather than controlled attributions of the performance gap.

5.2 Plug-and-Play Generalization across Backbones

Table 3 evaluates whether the complete CIFT training-and-deployment framework transfers across encoder families. The framework improves the corresponding linear-backbone control for all four encoders on Celeb-DF and DiffSwap, showing that its benefit is not restricted to ConvNeXt-V2-B. These full-framework-versus-linear comparisons test backbone portability rather than individual component effects; the controlled ablations below provide that attribution.

Table 3: Plug-and-play transfer of the CIFT framework across four encoder families under FF++ (c23) training. **Baseline**: corresponding backbone with a linear classifier, target-side diffusion, and no donor-dependent supervision. **Ours**: the complete donor-grounded training and source-free inference design with the same backbone. AUC (↑); Δ denotes the absolute AUC gain over the corresponding baseline. CDF = Celeb-DF; DS = DiffSwap. **Bold**: best Ours result or gain per column.

Backbone	Variant	CDF	Δ	DS	Δ
EfficientNet-B4	Baseline	69.90	—	66.78	—
	Ours	88.40	$+18.50$	84.03	$+17.25$
ResNet-50	Baseline	58.19	—	60.50	—
	Ours	75.45	$+17.26$	70.00	$+9.50$
ViT-B/16	Baseline	65.98	—	61.86	—
	Ours	78.57	$+12.59$	74.74	$+12.88$
ConvNeXt-V2-B	Baseline	72.61	—	75.42	—
	Ours (CIFT)	**95.30**	$\mathbf{+22.69}$	**98.47**	$\mathbf{+23.05}$

5.3 Ablation Study

We ablate CIFT along its dependency structure: the deployed source-free architecture, donor-side diffusion supervision, explicit XID-Mamba relational modeling, and IGS. Tables 4–6 report Celeb-DF, DiffSwap, and in-domain FF++; cross-domain effects are read from the first two columns, while Avg. is a three-dataset within-protocol summary distinct from the five-benchmark Avg. in Table 2.

Source-free architecture. Table 4 first isolates the detector retained at deployment while holding target-side auxiliary diffusion supervision fixed and using no donor-dependent training signal. The backbone-linear baseline reaches 75.35 average AUC. The Global Head and SMC individually reach 83.17 and 85.60, while their combination reaches 89.12. The combined source-free detector therefore improves the baseline by $+13.77$ and the stronger individual branch by $+3.52$, supporting complementary global and spatial evidence.

Table 4: Source-free architecture ablation. All rows use the same target-side auxiliary diffusion supervision and no donor-dependent training signal; only the deployed classification architecture is varied. Δ is measured relative to Backbone + Linear. All reported values are AUC. Avg. denotes the mean over Celeb-DF, DiffSwap, and FF++. **Bold**: best result per column.

Variant	Global	SMC	Celeb-DF	DiffSwap	FF++	Avg.	Δ vs. Backbone
Backbone + Linear	✗	✗	72.61	75.42	78.03	75.35	–
Global Head	✓	✗	81.82	82.44	85.25	83.17	$+7.82$
SMC	✗	✓	84.57	86.05	86.18	85.60	$+10.25$
Global Head + SMC	✓	✓	**87.83**	**89.45**	**90.08**	**89.12**	$\mathbf{+13.77}$

Donor-grounded relational learning. Table 5 compares progressively richer training configurations while keeping the deployed Global Head+SMC inference graph fixed. SF Base retains

target-side auxiliary diffusion supervision but uses no donor-dependent training signal. Donor Control additionally activates the donor stream through donor-side diffusion supervision while keeping XID-Mamba and IGS disabled, increasing average AUC from 89.12 to 93.17. This transition therefore measures donor-side diffusion supervision beyond the target-only SF Base, rather than explicit relational modeling. Donor-Grounded XID-Mamba then activates donor–target relational modeling while retaining the same target- and donor-side diffusion supervision, increasing average AUC to 94.63. Table 6 subsequently evaluates IGS, while Appendix C.5 separately evaluates donor and diffusion effects and Table 7 tests donor correspondence.

Table 5: Donor-grounded relational-learning ablation. The deployed Global Head+SMC inference graph is identical across all rows. SF Base uses target-side auxiliary diffusion supervision and no donor-dependent training signal. Donor Control adds the donor stream through donor-side diffusion supervision, without XID-Mamba or IGS. Donor-Grounded XID-Mamba retains both diffusion streams and additionally activates explicit donor–target relational modeling. All donor-dependent and diffusion components are discarded at inference. Δ is measured relative to SF Base. All reported values are AUC; Avg. denotes the mean over Celeb-DF, DiffSwap, and FF++. **Bold**: best result per column.

Variant	Donor stream	XID-Mamba relation	Celeb-DF	DiffSwap	FF++	Avg.	Δ vs. SF Base
SF Base	✗	✗	87.83	89.45	90.08	89.12	–
Donor Control	✓	✗	90.58	94.17	94.77	93.17	+4.05
Donor-Grounded XID-Mamba	✓	✓	**92.20**	**94.83**	**96.85**	**94.63**	**+5.51**

Relational geometry. Table 6 then isolates how the XID-Mamba relation is structured. Starting from 94.63 without IGS, cosine and gap constraints reach 95.19 and 95.37, respectively. Joint geometry raises average AUC to 97.16, and type-conditioned margins further improve it to 97.92, with gains on all three datasets. The larger joint and type-conditioned gains provide the primary evidence for the full XID-Mamba +IGS formulation. Together, these controlled comparisons distinguish the contribution of the deployed architecture, donor-side diffusion supervision, explicit XID-Mamba modeling, joint IGS, and type-conditioned IGS margins under a shared training and model-selection protocol.

Table 6: Relational-geometry ablation with the donor-grounded XID-Mamba representation fixed. $\mathcal{L}_{\mathrm{IG}}$ constrains directional cosine geometry, whereas $\mathcal{L}_{\mathrm{gap}}$ constrains Euclidean gap magnitude. "Type-cond. margin" indicates whether the active geometric loss uses forgery-type-dependent rather than fixed margins; "–" denotes not applicable when IGS is disabled. The XID-Mamba architecture is otherwise held fixed across rows. Δ is measured relative to XID-Mamba without IGS. All reported values are AUC. Avg. denotes the mean over Celeb-DF, DiffSwap, and FF++. **Bold**: best result per column.

Variant	$\mathcal{L}_{\mathrm{IG}}$	$\mathcal{L}_{\mathrm{gap}}$	Type-cond. margin	Celeb-DF	DiffSwap	FF++	Avg.	Δ vs. XID-Mamba
XID-Mamba, no IGS	✗	✗	–	92.20	94.83	96.85	94.63	–
Cosine Geometry	✓	✗	✗	92.83	95.21	97.54	95.19	+0.57
Gap Geometry	✗	✓	✗	92.35	96.63	97.13	95.37	+0.74
Joint Geometry	✓	✓	✗	94.57	97.74	99.16	97.16	+2.53
Type-Conditioned IGS	✓	✓	✓	**95.30**	**98.47**	**99.98**	**97.92**	**+3.29**

Table 7 tests whether the privileged gain depends on the training-time correspondence between the reference and observed face, rather than merely exposing the model to an additional face. The target images are fixed while only the forged-sample reference assignment changes. Average AUC decreases from 97.92 with the true donor to 95.96 with a same-identity non-corresponding donor, 94.96 when the target itself is reused, 93.54 with a random donor, and 93.40 after shuffling. DiffSwap is additionally subject-disjoint from the FF++ training identities, where CIFT reaches 98.47 AUC. These complementary controls show that the true-donor result is not reproduced by direct reuse of the target, generic second-face exposure, shuffled pairing, or memorization of the FF++ training subjects.

They do not imply that the learned representation contains no identity information. Identity-related information is intentionally used as privileged training supervision; the supported claim is that structured correspondence contributes beyond these tested alternatives.

Table 7: Donor-correspondence controls with the full XID-Mamba +IGS configuration fixed. Only the forged-sample donor assignment changes; the genuine-sample donor construction and source-free inference remain fixed. Thus, the "Target as donor" row changes only forged-sample donor correspondence. Avg. is over Celeb-DF, DiffSwap, and FF++; Δ is relative to the true donor. **Bold**: best per metric column.

Training donor assignment	Celeb-DF		DiffSwap		FF++		Avg.		Δ AUC
	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	
True donor x_s	**95.30**	**12.05**	**98.47**	**5.91**	**99.98**	**0.64**	**97.92**	**6.20**	–
Same-ID, non-corresponding donor	94.42	12.88	94.11	10.24	99.36	1.62	95.96	8.25	-1.95
Target as donor ($x_s{=}x$)	93.81	13.55	92.04	11.54	99.02	2.79	94.96	9.29	-2.96
Random wrong donor	92.37	15.41	89.54	15.37	98.72	3.14	93.54	11.31	-4.37
Shuffled donor assignment	91.48	16.72	90.86	13.82	97.86	4.21	93.40	11.58	-4.52

5.4 Relational-Space Analysis

We examine the learned relational geometry independently of classifier logits using the analysis-time gap $\Delta = \|g_s - g_t\|_2$. On held-out data, Δ achieves 93.72% AUROC on Celeb-DF and 90.06% on WildDeepfake (Table 14), showing separation between genuine and forged diagnostic pairs under the protocol in Appendix C.4. This diagnostic is not used by the deployed source-free classifier. Grad-CAM Selvaraju et al. (2019) visualizations are provided only as supplementary qualitative views in Appendix C.6; they are not used as causal or mechanistic evidence.

6 Discussion and Conclusion

CIFT investigates whether donor correspondence can provide privileged training supervision for source-free face-forgery detection. Trained only on FF++ (c23), the full model achieves **94.35** average AUC and **12.49** average EER across five held-out benchmarks. Under the encoder- and diffusion-prior-aligned EfficientNet-B4 comparison, CIFT improves over DiffusionFake by $+4.49$ average AUC, showing that the gain persists under this alignment, while the stronger ConvNeXt-V2-B encoder also contributes to final performance; mechanism attribution therefore relies on the aligned comparison and matched within-model controls. Under the controlled three-dataset ablation protocol (Celeb-DF, DiffSwap, and FF++), average AUC increases from $89.12 \rightarrow 93.17 \rightarrow 94.63 \rightarrow 97.16 \rightarrow 97.92$ as donor-side diffusion supervision, XID-Mamba, joint IGS, and type-conditioned IGS margins are introduced. With target images fixed, replacing the true donor with same-identity non-corresponding, target-as-donor, random, or shuffled alternatives consistently reduces performance, supporting a contribution from structured donor correspondence beyond the tested donor alternatives. Cross-manipulation transfer, perturbation robustness, and the analysis-time Δ diagnostic provide complementary evidence about transfer, robustness, and learned relational structure; the diagnostic is representation-level only, and all donor-dependent relational and diffusion components are removed at inference.

Limitations and future work. Training requires a donor x_s and, for forged samples, relation type τ, restricting direct use to donor-traceable corpora. In FF++, genuine donors come from the same video whereas forged donors come from the corresponding source video, so clip-level nuisance differences may remain. The fixed-target controls therefore support correspondence-dependent gains beyond the tested donor alternatives, not nuisance-invariant features; neither the perturbation study nor the analysis-time Δ diagnostic is interpreted as causal evidence for a unique forensic cue. The identity-gap interpretation applies to swaps; reenactment uses the weaker driving–target relation of Section 3.1. Training-only donor and diffusion pathways add offline cost but no deployment-time module. Future work should study nuisance-matched donor construction and weaker auxiliary supervision.

Reproducibility Statement

The experimental protocol, datasets, source-free evaluation setting, and primary metrics are specified in Section 4.1. The model formulation and separation between training-time privileged supervision and source-free inference are described in Sections 3.1–3.4, with detailed deployment and training algorithms in Appendices A.3–A.5. Backbone configurations, preprocessing, donor-reference construction, augmentation, optimization, and training cost are documented in Appendix B.1; loss coefficients, relational margins, hyperparameter selection, and checkpoint selection are given in Appendix B.2. Controlled evaluation and donor-correspondence protocols are specified in Appendices B.3 and B.4. Additional cross-manipulation, robustness, complexity, relational-space, seed-stability, video-level, and contemporary-benchmark analyses are provided in Appendix C. Held-out benchmark evaluations use the specified frozen checkpoints without target-benchmark fine-tuning or model-selection unless explicitly stated. Anonymized code is available at `https://github.com/aiai-9/CIFT`

Ethics Statement

This work studies defensive face-forgery detection using existing research benchmarks and introduces no new human-subject data or identity annotations. Donor/reference frames and identity-related information are used only as training supervision; no subject identity is predicted, and all donor-dependent computation is removed at deployment. Because facial imagery is privacy sensitive and detectors can make errors, CIFT should not be used as the sole basis for high-stakes attribution or identity-related decisions. We report robustness limitations without introducing forgery-generation or evasion procedures.

Statement on AI Use

Generative AI tools were used to assist with manuscript organization, language editing, readability, and critical feedback on the presentation of the methodology and experiments. All experimental design decisions, implementations, reported numerical results, mathematical statements, citations, and scientific claims were reviewed and verified by the authors. The authors take full responsibility for the content of the submission.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, May 2010. ISSN 0885-6125. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-End Reconstruction-Classification Learning for Face Forgery Detection . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4103–4112, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.00408. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00408.

Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis, 2024. URL https://arxiv.org/abs/2403.18471.

François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017. URL https://arxiv.org/abs/1610.02357.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. URL https://arxiv.org/abs/2006.07397.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization, 2023. URL https://arxiv.org/abs/2210.14457.

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, 2022. URL https://arxiv.org/abs/2203.01318.

Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. Google AI Blog, 2019. URL https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone, 2025. URL https://arxiv.org/abs/2407.08083.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. URL https://arxiv.org/abs/1709.01507.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4490–4499, 2023. doi: 10.1109/CVPR52729.2023.00436.

Younghun Kim, Minsuk Jang, Myung-Joon Kwon, Wonjun Lee, and Changick Kim. Selfi: Selective fusion of identity for generalizable deepfake detection, 2025. URL https://arxiv.org/abs/2506.17592.

Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset, 2021. URL https://arxiv.org/abs/2103.10094.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020. URL https://arxiv.org/abs/1909.12962.

Yuezun Li, Delong Zhu, Xinjie Cui, and Siwei Lyu. Celeb-df++: A large-scale challenging video deepfake benchmark for generalizable forensics, 2025. URL https://arxiv.org/abs/2507.18015.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020. URL https://arxiv.org/abs/2007.09355.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces, 2018. URL https://arxiv.org/abs/1803.09179.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images, 2022. URL https://arxiv.org/abs/2204.08376.

Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *AAAI Conference on Artificial Intelligence*, 2021a. URL https://api.semanticscholar.org/CorpusID:235306361.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin L, and Rongrong Ji. Dual contrastive learning for general face forgery detection, 2021b. URL https://arxiv.org/abs/2112.13522.

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, 2024. URL https://arxiv.org/abs/2410.04372.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures, 2019. URL https://arxiv.org/abs/1904.12356.

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. URL http://jmlr.org/papers/v16/vapnik15b.html.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018. URL https://arxiv.org/abs/1807.06521.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders, 2023. URL https://arxiv.org/abs/2301.00808.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection, 2023. URL https://arxiv.org/abs/2304.13949.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Weiming Zhang, Dongdong Chen, and Nenghai Yu. Multi-attentional deepfake detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2185–2194, 2021. doi: 10.1109/CVPR46437.2021.00222.

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8568–8577, 2023. doi: 10.1109/CVPR52729.2023.00828.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024. URL https://arxiv.org/abs/2401.09417.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection, 2024. URL https://arxiv.org/abs/2101.01456.

A Additional Methodology Details

A.1 Notation and Acronym Glossary

Table 8 summarizes the notation used in Section 3 and the supporting methodology appendix.

Table 8: Notation and acronyms used in the CIFT formulation.

Symbol / Acronym	Meaning
Core components	
CIFT	Consistency-Invariant Forensic Training.
XID-Mamba	Cross-Identity Mamba; training-only donor–target relational module, with the cross-identity interpretation applying to swaps.
IGS	Identity Gap Supervision; structures donor–target relational geometry, with the identity-gap interpretation applying to swaps.
SMC	Spatial Mamba Classifier; deployed source-free spatial classifier.
SD	Stable Diffusion.
MHA	Multi-Head Attention.
Inputs and labels	
$x \in \mathbb{R}^{3\times H\times W}$	Observed face image.
x_s	Donor reference used only during training; the identity donor for swaps and the driving/source reference for reenactment.
$y \in \{0, 1\}$	Binary label: 0 genuine and 1 forged.
τ	Manipulation type used for forged samples: `swap` or `reenact`.
Shared encoder and feature streams	
E	Shared image encoder; ConvNeXt-V2-B by default.
ϕ	Backbone feature projection.
$f_t,\ f_s$	Projected encoder features of x and x_s.
$U_t,\ U_s$	Independent target and donor stream transformations.
$\hat{f}_t,\ \hat{f}_s$	Transformed target and donor feature maps, Eq. 1.
$F_t,\ F_s$	Training-only FeatureFilter modules used by the diffusion pathway.
$\tilde{f}_t,\ \tilde{f}_s$	Filtered target and donor maps, Eq. 12.
Donor-grounded relational modeling	
$P_t,\ P_s$	Target and donor token projections before XID-Mamba.
$\tau_{\mathrm{sw}},\ \tau_{\mathrm{re}}$	Learnable swap and reenactment type tokens.
$\mathcal{R}_\tau$	Type-routing operation that prepends the selected token before the Mamba scan.
$\mathbf{Z}_t,\ \mathbf{Z}_s$	Target and donor token sequences after independent bidirectional Mamba encoding, Eq. 2.
$\bar{\mathbf{Z}}_t,\ \bar{\mathbf{Z}}_s$	Symmetrically cross-attended donor and target sequences, Eq. 3.
$g_t,\ g_s \in \mathbb{R}^{512}$	Mean-pooled target and donor relational embeddings, Eq. 4.
Δ	Type-conditioned donor-grounded relational gap $\|g_s - g_t\|_2$; an identity gap for swaps and a weaker driving–target discrepancy for reenactment.
Donor-grounded relational supervision	
$d_{\cos}$	Cosine distance $1 - \bar{g}_t{}^\top \bar{g}_s$.
δ	Genuine-pair pull tolerance, $\delta = 0.3$.
$m(\tau)$	Directional push margin: 1.0 for `swap`, 0.8 for `reenact`.
$m'(\tau)$	Euclidean push margin: 1.2 for `swap`, 0.9 for `reenact`.
$\mathcal{L}_{\mathrm{IG}}$	Directional cosine donor–target relational objective, Eq. 5.
$\mathcal{L}_{\mathrm{gap}}$	Euclidean donor–target gap objective, Eq. 6.
Deployed source-free classifier	
$\hat{y}_g$	Global Head classification logit.
$\mathbf{X} \in \mathbb{R}^{49\times 512}$	Spatial token sequence used by SMC.
α	Learned forward/reverse Mamba fusion coefficient $\sigma(\alpha_0)$.
$\hat{y}_s$	SMC classification logit, Eq. 11.
$\bar{w}_1, \bar{w}_2$	Learned source-free fusion weights.
$\bar{\beta}$	Source-free logit-fusion temperature.
$\hat{y}_{\mathrm{sf}}$	Final source-free prediction, Eq. 7.
Training-only auxiliary supervision	
U'_E	Trainable SD U-Net encoder copy used for conditioning.
$\mathcal{Z}$	Zero-initialized ControlNet convolutions Zhang et al. (2023).
$w_t,\ w_s$	Learned target and donor diffusion-injection coefficients $\sigma(\ell_{t,s})$.
$z_t^\eta,\ z_s^\eta$	Noisy VAE latents of the observed and donor images.
$\mathcal{L}_t,\ \mathcal{L}_s$	Target and donor denoising objectives, Eq. 13.
$\mathcal{L}_{\mathrm{diff}}$	Combined auxiliary diffusion loss $\mathcal{L}_t + \mathcal{L}_s$.
$\mathcal{L}_{\mathrm{reg}}$	Cross-attention regularization, Eq. 14.
$\mathcal{L}$	Complete training objective, Eq. 8.

A.2 Theoretical Motivation

The domain-adaptation bound Ben-David et al. (2010)

$$\varepsilon_{\mathcal{T}}(h) \le \varepsilon_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda^* \tag{9}$$

provides a useful lens for cross-pipeline generalization. The bound relates target error to source error, divergence between source and target domains, and the error of the best joint hypothesis. Features that are strongly coupled to a particular synthesis process can contribute to this domain divergence when the rendering pipeline changes.

CIFT does not optimize $d_{\mathcal{H}\Delta\mathcal{H}}$ directly. Instead, it uses donor-reference information to supervise a relation defined by the manipulation structure and then removes that privileged information at inference. The cross-dataset experiments test whether this training signal transfers better across pipelines. We therefore use Eq. 9 only as motivation, not as a quantitative guarantee.

A.3 Source-Free Classifier and Auxiliary Training Path

Global Head. The global branch applies average and max pooling to the observed-face feature map:

$$\hat{y}_g = C_g([\mathrm{GAP}(f_t); \mathrm{GMP}(f_t)])\,. \tag{10}$$

Spatial Mamba Classifier. The spatial branch pools f_t to a 7×7 grid and projects the result to $\mathbf{X} \in \mathbb{R}^{49\times512}$. Forward and reverse selective scans are combined using a learned coefficient $\alpha = \sigma(\alpha_0)$:

$$\begin{aligned} \mathbf{H} &= \alpha\,\mathrm{Mamba}(\mathbf{X}) + (1-\alpha)\,\mathrm{flip}(\mathrm{Mamba}(\mathrm{flip}(\mathbf{X})))\,, \\ \hat{y}_s &= C_s(\mathrm{mean}(\mathrm{LN}(\mathbf{H})))\,. \end{aligned} \tag{11}$$

The Global Head and SMC are the only classification branches retained at deployment.

Training-time classification. Eq. 7 is the sole binary classifier during both training and inference. XID-Mamba and IGS contribute only auxiliary relational gradients through the shared representation and never produce a classification logit.

Auxiliary diffusion regularization. For diffusion conditioning, independent FeatureFilter modules F_t and F_s transform the stream maps:

$$\tilde{f}_t = F_t(\hat{f}_t), \qquad \tilde{f}_s = F_s(\hat{f}_s). \tag{12}$$

A frozen Stable Diffusion 1.5 backbone Rombach et al. (2022) provides per-stream denoising supervision. A trainable U-Net encoder copy U'_E and zero-initialized ControlNet convolutions $\mathcal{Z}$ Zhang et al. (2023) inject the stream representation:

$$\mathcal{L}_t = \mathbb{E}_{\varepsilon,\eta}\left\|\varepsilon - \varepsilon_\theta\left(z_t^\eta, \eta, w_t\,\mathcal{Z}(U'_E(\tilde{f}_t))\right)\right\|_2^2, \tag{13}$$

with an analogous donor loss $\mathcal{L}_s$. Here z_t^η and z_s^η are noisy VAE latents and $w_t, w_s = \sigma(\ell_{t,s})$ are learned injection coefficients. The base SD U-Net, VAE, and CLIP encoder remain frozen; gradients update the trainable conditioning path and FeatureFilter modules. For the full donor-enabled model, the auxiliary objective is $\mathcal{L}_{\mathrm{diff}} = \mathcal{L}_t + \mathcal{L}_s$. In the progressive source-free controls, only the target-side term $\mathcal{L}_t$ is active, while Donor Control additionally activates the donor-side term $\mathcal{L}_s$. Appendix C.5 reports a separate factorial control of donor and diffusion effects within the full-model training configuration.

The remaining regularization term in Eq. 8 is

$$\mathcal{L}_{\mathrm{reg}} = \lambda_a \|\mathbf{A}_{\log}\|_1, \tag{14}$$

where $\mathbf{A}_{\log}$ denotes the cross-attention logits.

A.4 Type Conditioning

FF++ contains both identity-swapping and reenactment manipulations. Because these families do not impose the same relational structure, CIFT conditions XID-Mamba and IGS on manipulation type rather than using a single margin for all forged samples.

For forged samples, the selected type token is prepended to each pooled stream before the bidirectional Mamba scan:

$$\tau_{\mathrm{sw}}, \tau_{\mathrm{re}} \in \mathbb{R}^{1\times1\times512}, \qquad \tau_{\mathrm{sw}}, \tau_{\mathrm{re}} \sim \mathcal{N}(0, 0.02). \tag{15}$$

The token is removed after contextualization before cross-attention. Prefix placement is the default used in the main model; alternative conditioning mechanisms are evaluated in Table 16.

Table 9: FF++ manipulation families and relational supervision. Genuine samples use a distinct same-identity reference and no manipulation token.

FF++ family	Type	Relational regime
Deepfakes, FaceSwap	Swap	Stronger identity-gap margin
Face2Face, NeuralTextures	Reenactment	Reduced relational margin
YouTube genuine	Genuine	Pull same-identity pair together

A.5 Training and Inference Algorithm

Algorithm 1 summarizes the dependency structure of CIFT: the source-free detector is evaluated for every sample, while donor-based relational and diffusion supervision is activated only during training.

Algorithm 1 CIFT training and source-free inference

Require: Observed face x; during training also label y and donor reference x_s, and relation type τ for forged samples

$f_t \leftarrow \phi(E(x))$
Compute $\hat{y}_g, \hat{y}_s$ using Eqs. 10 and 11
Compute $\hat{y}_{\mathrm{sf}}$ using Eq. 7
if training **then**
    $f_s \leftarrow \phi(E(x_s))$
    Compute $\hat{f}_t, \hat{f}_s$ using Eq. 1
    Compute $\mathcal{L}_{\mathrm{diff}}$ using Eq. 13
    Compute g_t, g_s, Δ using Eqs. 2–4
    Compute $\mathcal{L}_{\mathrm{IG}}$ and $\mathcal{L}_{\mathrm{gap}}$ using Eqs. 5–6
    Optimize $\mathcal{L}$ using Eq. 8
else
    return $\hat{y}_{\mathrm{sf}}$
end if

A.6 Comparison Scope and Protocol Notes

Comparison scope. Our primary setting is source-free, frame-level visual detection. Because the methods in Table 2 may differ in preprocessing, frame sampling, input resolution, and capacity, the table provides benchmark context, while controlled attribution rests on the encoder-aligned comparison and the matched ablations. Methods using temporal information or different training formulations are not ranked directly, and Appendix C.9 examines only fixed score aggregation.

DiffusionFake comparison. DiffusionFake derives its source-related and target-related streams from the encoded forgery, whereas CIFT separately encodes the donor reference as a relational training input.

Substituting ConvNeXt-V2-B into DiffusionFake would alter its reference architecture. We therefore use EfficientNet-B4 + Ours as the closest controlled comparison: it retains the encoder family and frozen SD 1.5 prior of DiffusionFake while adding donor-grounded XID-Mamba/IGS supervision. The full ConvNeXt-V2-B configuration is reported separately as CIFT.

B Additional Experimental Details

B.1 Implementation Details

Backbone and input. The default encoder is ConvNeXt-V2-Base Woo et al. (2023) (`convnextv2_base.fcmae_ft_in22k_in1k` from timm). Face crops are resized to 256×256, and the stage-3 output is projected to $D = 1792$ through ϕ. For the backbone-transfer experiments in Section 5.2, EfficientNet-B4, ResNet-50, and ViT-B/16 replace ConvNeXt-V2-B and use 224×224 inputs.

Stream transformations. The target and donor streams use independent transformations U_t and U_s. Each applies spatial upsampling before the features are used by XID-Mamba and the auxiliary diffusion pathway. For diffusion conditioning, the corresponding FeatureFilter F_t or F_s comprises a 3×3 convolution with SiLU, SE channel attention Hu et al. (2019), CBAM spatial attention Woo et al. (2018), 8-head self-attention on a 7×7 token grid, and a 1×1 projection to $C' = 320$. Target and donor filters do not share parameters.

SMC and XID-Mamba. Both modules use token dimension $d = 512$ and spatial pooling to a 7×7 grid. SMC contains two bidirectional Mamba layers Gu & Dao (2024) with state dimension 16, dropout 0.1, and DropPath 0.1. XID-Mamba uses one bidirectional Mamba layer per stream followed by 8-head symmetric cross-attention. Its outputs are used only by the training-time relational objectives.

Auxiliary diffusion pathway. The diffusion pathway is initialized from Stable Diffusion 1.5 Rombach et al. (2022). A trainable encoder copy and zero-initialized ControlNet convolutions Zhang et al. (2023) inject stream features into the denoising pathway. The base SD U-Net, VAE, and CLIP encoder remain frozen. Learned injection coefficients $w_t, w_s = \sigma(\ell_{t,s})$ control conditioning strength at 13 resolution levels per stream. This pathway is used only during training and contributes no module or diffusion computation to the deployed source-free detector.

Optimization. We use AdamW Loshchilov & Hutter (2019) with base learning rate $\eta = 2\times10^{-5}$ and weight decay 5×10^{-4}. The classification heads, source-free fusion, XID-Mamba, and relational-objective parameters use learning rate 1.2η. The per-GPU micro-batch size is 32, and gradients are accumulated for four steps before each optimizer update. OneCycleLR uses an 8% warm-up followed by cosine annealing. Gradients are clipped to norm 0.5, and all models are trained for 80 epochs using bf16 distributed data parallelism on $4\times$ NVIDIA A100 80 GB GPUs. The full model requires approximately 12 hours of wall-clock training.

Data augmentation. Training augmentations include JPEG compression with quality sampled from $[30, 100]$, Gaussian blur, additive noise, colour jitter, greyscale, random erasing, and SBI augmentation Shiohara & Yamasaki (2022) for genuine samples.

Donor-reference retrieval. For each forged FF++ frame, x_s is the temporally nearest valid frame from the corresponding source video passing an MTCNN landmark-confidence threshold of 0.95. For genuine samples, we exclude the target frame and select the nearest distinct valid frame from the same video. Up to 50 frames per video are sampled using the predefined FF++ splits.

B.2 Training Configuration

The loss coefficients used for all reported experiments are

$$\begin{aligned} \lambda_{\text{cls}} = 1.5, \quad \lambda_{\text{diff}} = 0.5, \quad \lambda_{\text{ig}} = 0.15, \\ \lambda_{\text{gap}} = 0.1, \quad \lambda_a = 10^{-4}. \end{aligned} \tag{16}$$

The binary classification objective is $\mathcal{L}_{\text{cls}} = 0.8\,\mathcal{L}_{\text{BCE}} + 0.2\,\mathcal{L}_{\text{focal}}$ with focal parameters $\gamma = 2.0$ and $\alpha_f = 0.25$ Lin et al. (2018).

The IGS configuration uses genuine-pair tolerance $\delta = 0.3$. For the non-type-conditioned geometry controls in Table 6, all forged samples use fixed margins $m = 1.0$ and $m' = 1.2$. The full type-conditioned setting instead uses $m(\tau) = 1.0/0.8$ and $m'(\tau) = 1.2/0.9$ for swap/reenactment, respectively. The source-free fusion parameters $\bar{w}_1, \bar{w}_2, \bar{\beta}$ are optimized jointly with the deployed classifier. Loss weights and IGS margins were fixed using only FF++ training/validation. The Mamba relational head and prefix type-token placement were specified before held-out evaluation and were not selected by comparing variants on Celeb-DF or DiffSwap. The alternative relational-head and type-token results in Tables 12 and 16 are post-selection sensitivity controls. Every run selects the checkpoint with highest FF++ validation AUC, and controlled variants inherit the same optimization, schedule, training budget, and selection criterion unless the varied factor defines the comparison.

B.3 Controlled Evaluation Protocol

The main cross-dataset evaluation reports AUC and EER on five held-out benchmarks and uses no donor input at inference. The controlled ablations in Tables 4– 7 instead use Celeb-DF, DiffSwap,

and FF++ so that successive training interventions can be compared under a common setting. Their reported average is therefore a three-dataset mean and is not directly comparable with the five-benchmark average in Table 2.

Source-free architecture control. The backbone with a linear classifier, Global Head, SMC, and their combination use target-side auxiliary diffusion supervision $\mathcal{L}_t$ but no donor-dependent training signal. Only the deployed classification architecture changes across these rows.

Donor and relational controls. The deployment graph and classification objective are fixed across all variants. Among the auxiliary training signals, SF Base uses only the target-side diffusion term $\mathcal{L}_t$ and no donor-dependent signal. Donor Control additionally activates the donor stream and donor-side diffusion loss $\mathcal{L}_s$, while XID-Mamba, $\mathcal{L}_{\mathrm{IG}}$, and $\mathcal{L}_{\mathrm{gap}}$ remain disabled. The XID-Mamba variant retains both diffusion terms and activates explicit donor–target relational modeling; Table 6 then introduces the IGS objectives.

Geometry controls. With XID-Mamba fixed, we evaluate directional supervision $\mathcal{L}_{\mathrm{IG}}$, magnitude supervision $\mathcal{L}_{\mathrm{gap}}$, their joint formulation, and type-conditioned margins. These settings correspond directly to Table 6.

Backbone controls. EfficientNet-B4, ResNet-50, ViT-B/16, and ConvNeXt-V2-B are evaluated with corresponding linear controls and complete CIFT variants under the same FF++ protocol. Every linear control uses target-side diffusion $\mathcal{L}_t$ and no donor-dependent supervision. The ConvNeXt-V2-B control is therefore the same Backbone + Linear configuration reported in Table 4. EfficientNet-B4 additionally provides the closest encoder- and diffusion-prior-aligned comparison with Diffusion-Fake Sun et al. (2024).

B.4 Donor Correspondence and Memorization Controls

Because CIFT deliberately uses identity-related reference information during training, the learned representation may retain subject information; identity invariance is neither assumed nor claimed. Our evaluation instead tests whether simpler alternatives can account for the donor-dependent gain: memorization of FF++ training subjects, reuse of the target itself, generic exposure to another face, or donor assignments that preserve identity or image marginals while weakening sample-specific correspondence. No subject-identity label is provided to the model or used as a training target; identity grouping is used only to construct the controlled same-identity reference assignment. All donor-dependent computation is removed at deployment.

True donor. The source frame associated with the forged target is used as x_s. This is the full CIFT configuration.

Same-identity, non-corresponding donor. For forged samples, the true donor is replaced by a different frame from the same source identity; genuine-sample donor construction is unchanged.

Target as donor. For forged samples, we set $x_s = x$ while preserving the two-stream computation; genuine-sample donor construction is unchanged.

Random wrong donor. For forged samples, the donor is replaced by an unrelated source identity; genuine-sample donor construction is unchanged.

Shuffled donor assignment. Donors are randomly permuted among forged samples within each training batch; genuine-sample donor construction is unchanged. This preserves the forged-sample donor distribution while disrupting sample-specific donor–target correspondence.

Table 7 reports donor-correspondence controls in which the forged targets, optimization, and deployed inference graph remain fixed while only donor assignment changes. These interventions test whether the true donor contributes beyond identity-preserving non-correspondence, the target itself, unrelated donors, and shuffled assignment. Because the genuine-donor construction remains fixed, they do not claim to remove all clip-level nuisance information.

Source-free evaluation and interpretation. All external benchmarks are evaluated from the observed face alone; no reference image, source-video identifier, subject label, or correspondence metadata is available to the deployed classifier. The 6,000-image DiffSwap partition is subject-disjoint from the FF++ training identities, so its result cannot be explained by direct recognition of those training subjects. Separately, performance decreases when the true donor is replaced by a

same-identity non-corresponding donor, the target itself, a random donor, or a shuffled assignment (Table 7). These interventions preserve different aspects of identity or second-image information while disrupting the intended correspondence in different ways. Together, these controls support the narrower conclusion that structured donor correspondence contributes information not reproduced by the tested memorization and donor alternatives. They do not establish that clip-level nuisance information is absent, and CIFT does not require identity-invariant internal representations.

C Additional Results and Ablations

This appendix complements the main cross-dataset results and controlled ablations with additional tests of cross-manipulation transfer, perturbation robustness, relational-head design, computational complexity, auxiliary diffusion supervision, type-token injection, seed stability, and video-level aggregation. Unless stated otherwise, all models are trained on FF++ (c23) and use source-free inference.

C.1 Cross-Manipulation Generalization

Table 10: Cross-manipulation evaluation on FF++ (c23) in terms of AUC (↑). Rows give the training manipulation and columns the test manipulation; diagonal entries are in-domain. EfficientNet-B4 + Ours shares the encoder family and frozen SD 1.5 prior of DiffusionFake. **Bold**: best off-diagonal result in each transfer setting.

Train	Method	DF	F2F	FS	NT
DF	EfficientNet-B4 + Ours	99.98	77.41	52.33	75.28
DF	DiffusionFake	99.82	78.46	52.29	74.43
DF	CIFT	99.96	**84.46**	**65.38**	**81.55**
F2F	EfficientNet-B4 + Ours	85.14	99.54	61.98	64.86
F2F	DiffusionFake	88.92	99.36	63.19	68.55
F2F	CIFT	**94.57**	99.61	**79.45**	**82.95**
FS	EfficientNet-B4 + Ours	76.44	69.86	99.32	52.84
FS	DiffusionFake	75.28	70.91	99.12	52.17
FS	CIFT	**79.88**	**76.18**	99.51	**59.86**
NT	EfficientNet-B4 + Ours	90.79	52.82	78.57	98.42
NT	DiffusionFake	89.54	51.71	79.15	98.71
NT	CIFT	**93.76**	**65.39**	**84.84**	99.28

Table 10 evaluates transfer when training is restricted to one FF++ manipulation family. CIFT obtains the highest AUC in all 12 off-diagonal settings while maintaining at least 99.28 AUC on the corresponding in-domain manipulations. The largest improvements over DiffusionFake arise for F2F→FS (63.19 → 79.45, +16.26) and F2F→NT (68.55 → 82.95, +14.40). The EfficientNet-B4 variant performs comparably to DiffusionFake, exceeding it in 6 of 12 off-diagonal settings by at most 1.25 AUC.

C.2 Robustness to Image Perturbations

Table 11: Robustness to image perturbations on Celeb-DF (AUC↑). All models are trained on FF++ (c23). Perturbations follow the CADDM Dong et al. (2023) protocol at moderate severity. Samples for which perturbation causes face-detection failure (MTCNN confidence < 0.95) are excluded. **Bold**: best result per column.

Method	JPEG	Blur	Noise	Sat.	Cont.	Block	Avg.
Xception	61.3	58.4	60.1	64.8	63.9	55.2	60.6
CADDM	74.2	76.8	75.3	78.1	77.4	73.9	76.0
DiffusionFake	80.7	81.4	79.8	84.1	83.5	80.9	81.7
CIFT	**91.8**	**86.2**	**88.7**	**92.4**	**88.3**	**87.9**	**89.2**

Across the six perturbations, CIFT obtains the highest AUC in every setting and averages 89.2, compared with 81.7 for DiffusionFake and 76.0 for CADDM. These results establish robustness

under the evaluated post-processing operations. Because JPEG, blur, and noise overlap with training augmentations, we do not interpret this experiment as an isolated causal test of frequency reliance; Table 7 separately tests sensitivity to donor assignment.

C.3 Relational-Head and Complexity Controls

We next test whether the effect of the donor reference depends on the training-time relational operator. The deployed source-free graph, training schedule, and evaluation protocol remain fixed.

Table 12: Relational-head control on Celeb-DF and DiffSwap. Within each head, $x_s{=}x$ for every training sample removes the distinct donor input while preserving the same two-stream interface; the true-donor row restores the default training-time donor construction. "Gain" is the AUC change from donor removal to the true donor within the same relational head. This differs from Table 7, where only the forged-sample donor assignment is changed while genuine-sample donor construction remains fixed.

Head	Training input	Celeb-DF			DiffSwap		
		AUC↑	EER↓	Gain	AUC↑	EER↓	Gain
MLP	$x_s = x$	81.14	26.83	–	86.27	19.54	–
	true x_s	83.49	24.17	+2.35	88.61	17.82	+2.34
Transformer	$x_s = x$	94.53	13.02	–	93.58	10.26	–
	true x_s	91.86	16.07	-2.67	90.52	17.34	-3.06
Mamba	$x_s = x$	94.50	12.66	–	93.45	10.30	–
	true x_s	**95.30**	**12.05**	+0.80	**98.47**	**5.91**	+5.02

The MLP and Mamba relational heads improve when the true donor is used, whereas the Transformer control decreases on both benchmarks. Donor access therefore does not guarantee improved transfer for an arbitrary relational operator. Among the tested heads, Mamba achieves the strongest absolute performance and benefits most from the true donor on DiffSwap.

Table 13: Complexity of the relational-head controls. All variants share the same deployed source-free graph. FLOPs correspond to one training forward pass at $256{\times}256$.

Metric	MLP	Transformer	Mamba/XID-Mamba
Inference parameters	103.4M	103.4M	103.4M
Training parameters	421.42M	428.78M	423.47M
Training FLOPs	484.115G	484.709G	484.331G
Relative FLOPs vs. MLP	–	$+0.12\%$	$+0.04\%$

All three controls use 103.4M parameters at inference because the training-time relational head is removed before deployment. Mamba/XID-Mamba increases training forward FLOPs by only 0.04% relative to the MLP control and remains below the Transformer configuration. The performance differences in Table 12 therefore cannot be explained by deployed parameter count or forward FLOPs alone.

C.4 Analysis-Time Relational-Gap Diagnostic

The main experiments evaluate the deployed source-free classifier. Here we perform a separate representation-level diagnostic to examine whether the donor-grounded geometry learned by XID-Mamba and IGS remains structured on held-out data. This analysis does not modify the deployed inference protocol and is not used to produce any result in Table 2.

Diagnostic protocol. For this analysis only, the frozen training-time XID-Mamba branch is reactivated using fixed same-identity query–reference pairs. Pair construction is fixed before computing classifier logits or Δ scores, and the exact query and reference file identifiers for every evaluated pair are released with the diagnostic code. These references are used only for representation analysis; the deployed detector receives no reference image, does not execute XID-Mamba, and does not use Δ for classification.

For each diagnostic pair, we compute

$$\Delta = \|g_s - g_t\|_2$$

without using classifier logits. Table 14 reports the resulting distribution statistics.

Table 14: Relational gap Δ under the analysis-time diagnostic. The reference image is used only to probe the frozen relational space and is never available to the deployed classifier. $\text{Sep} = \mu_f - \mu_r$ denotes the difference between forged and genuine mean gaps.

Dataset	N_r	N_f	μ_r	σ_r	μ_f	σ_f	Sep	AUROC	EER
FF++	1000	4994	0.060	0.035	1.642	0.061	1.582	99.98	0.07
Celeb-DF	877	5621	0.343	0.511	1.610	0.242	1.266	93.72	8.40
DFD	241	2578	0.422	0.578	1.561	0.349	1.140	88.90	14.07
WildDeepfake	2340	3091	0.233	0.447	1.293	0.621	1.060	90.06	13.29

Across the four evaluated datasets, the forged and genuine diagnostic distributions remain separated in the learned relational space. The gap alone reaches 99.98 AUROC on FF++, 93.72 on Celeb-DF, 90.06 on WildDeepfake, and 88.90 on DFD. Because this analysis introduces an auxiliary reference that is absent from deployment, these values are interpreted only as evidence about the geometry learned during training, not as source-free detection results.

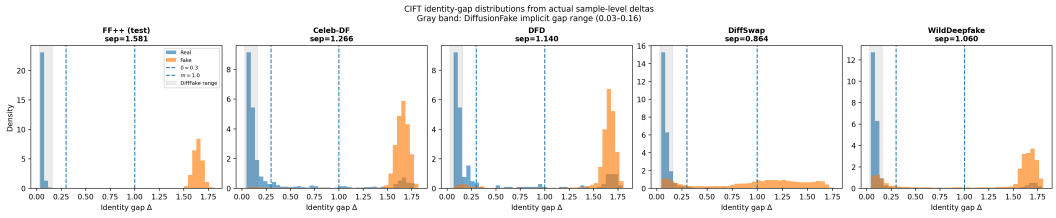

Figure 2: Distribution of analysis-time relational gaps Δ across four datasets. Genuine and forged samples occupy different regions of the relational space. Dashed lines show the genuine pull tolerance $\delta = 0.3$ and the swap directional margin $m = 1.0$ used during training. These distributions are diagnostic only and are not used by the deployed classifier.

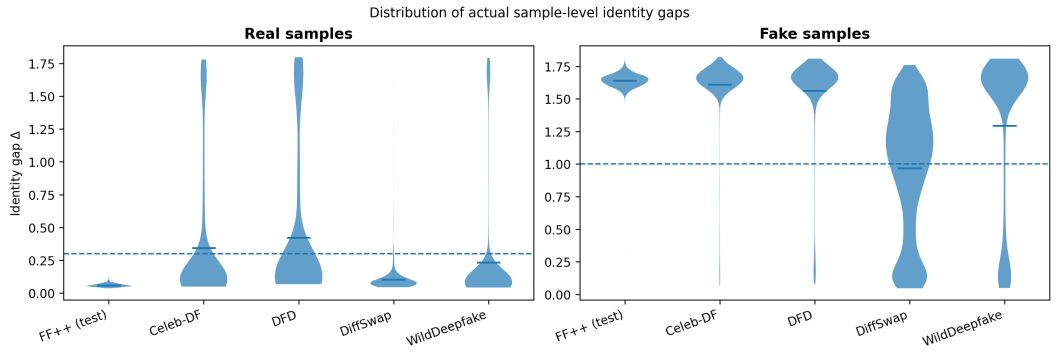

Figure 3: Violin distributions of the analysis-time relational gap Δ for genuine and forged samples. Separation is strongest on FF++ and Celeb-DF and remains visible on WildDeepfake and DFD. The diagnostic reference is not available during source-free inference.

C.5 Auxiliary Training Signal Ablation

Within this factorial control, removing diffusion lowers average AUC from 97.92 to 95.10, replacing the distinct donor with the target for all samples lowers it to 95.56, and applying both interventions lowers it to 92.48. On diffusion-generated DiffSwap, removing diffusion retains 97.29 AUC versus

Table 15: Factorial control of donor and auxiliary diffusion effects within the full-model training configuration. "Donor" indicates whether a distinct training-time donor is supplied. The "Full − donor" row sets $x_s{=}x$ for every training sample and uses the same Mamba donor-removed configuration reported in Table 12; the identical values are therefore intentional. This differs from the "Target as donor" intervention in Table 7, where only the forged-sample donor assignment changes while genuine-sample donor construction remains fixed. "Diffusion" indicates whether the auxiliary diffusion pathway is active. All settings share the same backbone, training schedule, model-selection rule, and source-free inference graph. Avg. is the mean over Celeb-DF, DiffSwap, and FF++. **Bold**: best result per metric column.

Setting	Donor	Diffusion	Celeb-DF		DiffSwap		FF++		Avg.	
			AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
Full model	✓	✓	**95.30**	**12.05**	**98.47**	**5.91**	**99.98**	**0.64**	**97.92**	**6.20**
Full − diffusion	✓	✗	90.23	16.18	97.29	7.23	97.79	7.17	95.10	10.19
Full − donor	✗	✓	94.50	12.66	93.45	10.30	98.73	3.88	95.56	8.95
Full − diffusion − donor	✗	✗	89.96	16.29	91.55	15.91	95.94	11.86	92.48	14.69

98.47 for the full model, showing that auxiliary diffusion contributes but does not account for the result alone.

This factorial study is separate from the progressive main-text ablations and from the forged-sample donor-assignment controls in Table 7. The "Full − donor" Mamba condition is the donor-removed $x_s{=}x$ configuration also used in Table 12; it is reused here to complete the donor×diffusion factorial. Table 7 instead tests donor correspondence by changing only the forged-sample donor assignment. The resulting deltas therefore quantify different donor interventions and are not alternative estimates of the same effect.

C.6 Qualitative Grad-CAM Visualization

Figure 4 provides supplementary Grad-CAM Selvaraju et al. (2019) visualizations for the deployed source-free classifier on representative genuine and forged examples from Celeb-DF and DiffSwap. Grad-CAM is used only to visualize prediction-specific spatial saliency and is not part of training, inference, model selection, or quantitative evaluation.

In the displayed examples, genuine predictions exhibit relatively diffuse saliency, whereas several forged examples show stronger localized responses around central facial regions. We report these observations descriptively only: Grad-CAM can vary across samples and does not establish feature necessity or causal dependence. Accordingly, no claim about the detector's forensic mechanism is derived from Figure 4. Evidence for the proposed training mechanism instead comes from the controlled architecture, relational-geometry, and donor-correspondence interventions in Tables 4–7.

C.7 Type-Token Injection Control

Prefix injection performs best among the tested mechanisms on both benchmarks. Relative to removing the XID-Mamba type token, it improves AUC by 4.08 points on Celeb-DF and 1.76 points on DiffSwap and reduces EER by 4.34 and 3.37 points, respectively. These post-selection results are consistent with the prefix-token configuration used by CIFT and characterize sensitivity to type injection without implying a causal effect of Mamba scan order.

C.8 Seed Stability

The standard deviation of CIFT does not exceed 0.18 AUC across runs, well below its margin over DiffusionFake on both benchmarks. Table 2 reports the default-seed run of CIFT, fixed before held-out evaluation. These runs characterize variability only, and no significance test is performed.

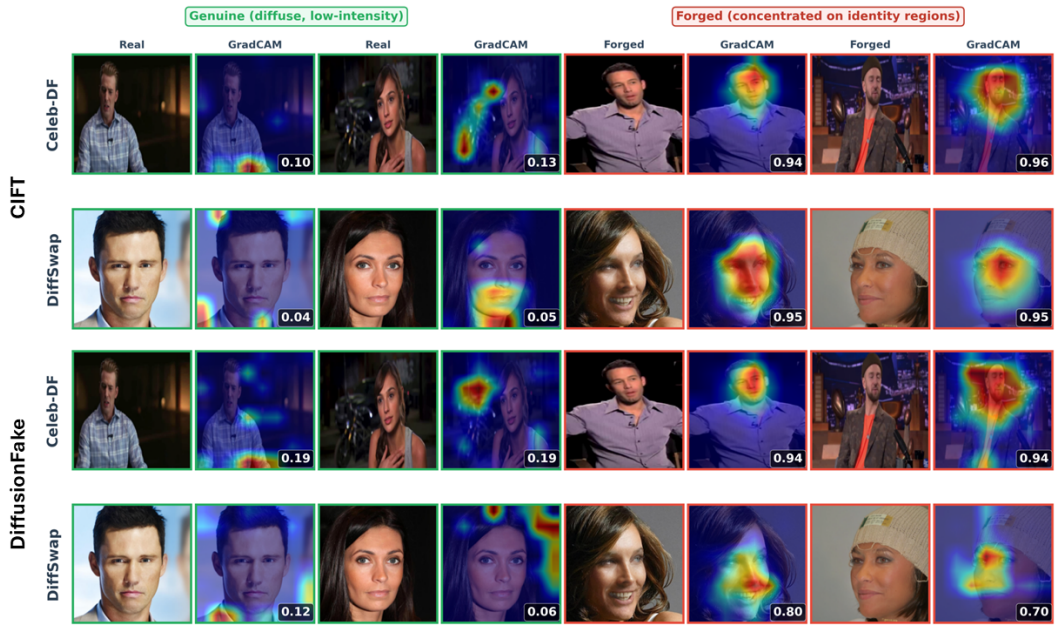

Figure 4: Qualitative Grad-CAM Selvaraju et al. (2019) visualizations for CIFT (rows 1–2) and DiffusionFake Sun et al. (2024) (rows 3–4) on Celeb-DF and DiffSwap. The heatmaps visualize gradient-based spatial saliency for each prediction; detection scores are shown in the bottom-right of each example. These examples are included only to illustrate prediction-specific activation patterns and are not interpreted as identifying causal features, stable forensic regions, or the mechanism responsible for the reported performance.

Table 16: Type-token injection in XID-Mamba. The backbone, XID-Mamba architecture, IGS objectives and margins, training schedule, and source-free inference protocol are fixed; only the mechanism used to inject type information into XID-Mamba is varied. **Bold**: best result per metric column.

Conditioning	Celeb-DF		DiffSwap	
	AUC↑	EER↓	AUC↑	EER↓
No XID-Mamba type token	91.22	16.39	96.71	9.28
FiLM Perez et al. (2017)	92.04	16.01	93.83	14.73
Suffix token (pos. $N+1$)	92.77	15.75	94.67	13.98
Prefix token (pos. 0, CIFT)	**95.30**	**12.05**	**98.47**	**5.91**

C.9 Video-Level Evaluation

The deployed CIFT model remains a frame-level detector. For this evaluation, frame logits are mean pooled within each video without temporal modeling. CIFT reaches 96.81 AUC on Celeb-DF and 93.47 on WildDeepfake, exceeding DiffusionFake by 11.38 and 15.65 points, respectively. Thus, the frame-level scores remain informative under fixed video-level aggregation, without implying a comparison to learned temporal detectors.

C.10 Contemporary Benchmark Evaluation

We further test whether the frozen FF++-trained detectors transfer to three additional benchmarks not used for training, validation, model selection, or hyperparameter tuning: KoDF Kwon et al. (2021), Celeb-DF++ Li et al. (2025), and DiffusionFace Chen et al. (2024). No fine-tuning, retuning,

Table 17: Seed stability on Celeb-DF and DiffSwap (AUC↑). Values are mean ± standard deviation over three independent runs with fixed splits and method-specific hyperparameters.

Method	Celeb-DF	DiffSwap
DiffusionFake Sun et al. (2024)	83.17±0.12	82.02±0.15
CIFT	**95.28±0.16**	**98.31±0.18**

Table 18: Video-level evaluation on Celeb-DF and WildDeepfake. Frame logits are mean pooled within each video before computing AUC and EER. No temporal model or additional tuning is used. Avg. is the mean over the two datasets. **Bold**: best result per column.

Method	Celeb-DF		WildDeepfake		Avg.	
	AUC↑	EER↓	AUC↑	EER↓	AUC↑	EER↓
DiffusionFake Sun et al. (2024)	85.43	22.17	77.82	31.04	81.63	26.61
CIFT	**96.81**	**10.22**	**93.47**	**13.08**	**95.14**	**11.65**

calibration, or test-set selection is performed. CIFT and DiffusionFake use identical face detections, crops, and sampled frames within each benchmark while retaining their native input resolution and normalization.

For KoDF and Celeb-DF++, both detectors are evaluated on the same sampled frames and frame logits are mean pooled within each video before computing AUC. DiffusionFace is evaluated per image. Because these datasets mix video-level and image-level evaluation, we report them separately rather than combining them with the five-benchmark average in Table 2.

Table 19: Zero-shot evaluation on additional contemporary benchmarks using frozen FF++-trained checkpoints. KoDF and Celeb-DF++ use mean-pooled video predictions; DiffusionFace is evaluated per image. No target-benchmark fine-tuning or model selection is performed. AUC (↑).

Dataset	Level	DiffusionFake	CIFT	Gain
KoDF	Video	79.28	**87.14**	+7.86
Celeb-DF++	Video	74.88	**86.75**	+11.87
DiffusionFace	Image	82.61	**89.36**	+6.75

Across all three additional benchmarks, CIFT improves over DiffusionFake without adaptation, with gains of +7.86, +11.87, and +6.75 AUC on KoDF, Celeb-DF++, and DiffusionFace, respectively. Together with the primary five-benchmark evaluation, these results provide additional evidence that the observed generalization is not restricted to the original FF++ evaluation suite.

The comparison remains within the source-free visual setting. Dedicated temporal and audio-visual detectors use additional information or training protocols and are therefore not ranked directly against these frame-based models. As shown in Appendix C.9, CIFT can additionally mean pool frame logits at video level without introducing temporal modeling.