RELATION CONFLICT ESTIMATION FOR GENERALIZABLE FACE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizable face forgery detection remains challenging because models trained on known manipulations often overfit to specific appearance artifacts, which may not transfer to unseen forgery mechanisms. However, the discriminative evidence carried by facial relation conflicts remains insufficiently explored. We observe that authentic facial videos tend to preserve stronger compatibility between facial patch representations and their surrounding contexts, whereas forgery synthesis may introduce subtle contextual discrepancies even when local appearances remain visually plausible. Motivated by this observation, we propose a Relation Conflict Estimation framework for generalizable face forgery detection, which exploits contextual discrepancies among facial regions as transferable forgery evidence. Lightweight distributional guidance is further introduced to regularize relation representations in terms of frame-level conflict magnitude and cross-frame distribution. Extensive experiments under cross-dataset, cross-manipulation, and cross-generation protocols demonstrate that the proposed method consistently improves generalization across both manipulation-based and AIGC-generated forgeries, with consistent gains observed on ViT-B/16 and CLIP ViT-L/14 backbones. Ablation and visualization analyses further show that pairwise relation-conflict separation provides the primary regularization effect, while cross-frame distribution guidance offers complementary improvements.

1 INTRODUCTION

Recent advances in DeepFake and generative models have enabled the creation of realistic forged facial images and videos, posing growing threats to digital media authenticity. Many face forgery detection methods Wang et al. (2023b); Yan et al. (2024a); Tan et al. (2024); Lai et al. (2025); Cui et al. (2025); Yan et al. (2025a) capture manipulation traces from local textures, blending boundaries, and frequency statistic features. In addition, some methods Choi et al. (2024); Zhang et al. (2024); Guo et al. (2025b); Yan et al. (2025b); Han et al. (2025) further exploit temporal evidence such as abnormal facial motion, physiological inconsistency, or discontinuities across adjacent frames. Although effective when characteristic artifacts are present, these cues are often tied to specific manipulation pipelines and can be weakened by improved generators, compression, or post-processing. As a result, detectors trained on known manipulations may overfit to manipulation-specific appearance patterns and generalize poorly to unseen forgery mechanisms.

Beyond directly observable artifacts, recent studies have further explored contextual dependencies among facial regions Bai et al. (2023); Huang et al. (2023); Xu et al. (2026), yet the discriminative evidence carried by contextual relation conflicts remains comparatively underexplored. Our key observation is that authentic facial videos tend to preserve stronger compatibility between facial patch representations and their surrounding contexts, whereas face swapping and AIGC-based synthesis may introduce subtle contextual discrepancies even when local appearances remain visually plausible. Such discrepancies describe how a facial region agrees with its broader context, and may therefore provide complementary evidence beyond isolated appearance artifacts.

Motivated by this observation, we propose a Relation Conflict Estimation framework for generalizable face forgery detection. The proposed method estimates context-induced relation responses for facial patch representations and leverages their discrepancies as relation-conflict evidence for forgery prediction. To further improve relation learning, we introduce lightweight distributional

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.

guidance that regularizes both frame-level conflict magnitude and cross-frame conflict distribution. In particular, authentic and forged samples are encouraged to maintain separable relation-conflict magnitudes, while authentic videos are further regularized to exhibit a stable cross-frame conflict distribution. These training objectives introduce no additional inference-time computation.

Extensive experiments under cross-dataset, cross-manipulation, and cross-generation protocols demonstrate that the proposed method consistently improves generalization across both manipulation-based and AIGC-generated forgeries. Consistent gains are observed with both ViT-B/16 and CLIP ViT-L/14 backbones, validating the effectiveness of the proposed relation-conflict modeling across backbone architectures of different scales.

Our main contributions are summarized as follows:

- We propose a Relation Conflict Estimation framework for generalizable face forgery detection, which characterizes contextual discrepancies between facial patch representations and their context-induced relation responses and incorporates the resulting relation evidence into forgery prediction.
- We introduce lightweight distributional guidance for relation learning, including pairwise relation-conflict separation and cross-frame distribution regularization, to improve the discriminability and generalization of the learned relation representations.
- Extensive experiments under cross-dataset, cross-manipulation, and cross-generation settings demonstrate consistent improvements across both manipulation-based and AIGC-generated forgeries, with ablation and visualization analyses further validating the proposed design.

2 Related Work

2.1 Forgery Evidence Modeling for Generalization

Existing face forgery detection methods exploit diverse forgery-related evidence, including local texture artifacts, blending boundaries, frequency statistics, and semantic representations Wang et al. (2025); Yan et al. (2025a); Jia et al. (2020); Zhao et al. (2021); Zhuang et al. (2022); Guo et al. (2025a). Conv-MLP combines convolutional inductive bias with MLP-based global modeling to jointly capture local forgery traces and global semantic information Wang et al. (2022a). Forensics Adapter introduces a lightweight forensic branch on top of CLIP to capture forgery-sensitive cues such as blending boundaries Cui et al. (2025), while GM-DF employs a Mixture-of-Experts framework together with CLIP-guided alignment to improve cross-scenario generalization Lai et al. (2025).

Some works aim to improve generalization by enlarging the diversity of training distributions. LSDA simulates latent variations among different forgery types to learn more transferable representations Yan et al. (2024a). NDA-FAS synthesizes negative samples from authentic faces through negative data augmentation Wang et al. (2023a), while style-transfer-based methods reduce domain discrepancy by transforming target-domain appearances toward the source domain Zhou et al. (2024); Wang et al. (2022b); Choi et al. (2024). Other approaches further generate diverse forgery samples to enrich training distributions Ge et al. (2024); Otake et al. (2024); Wang et al. (2023a).

In addition, some studies exploit frequency-aware or structural cues. SFDG combines frequency extraction with dynamic reasoning to complement spatial representations Wang et al. (2023b), while FreqNet adopts frequency-aware modeling to reduce dependence on generator-specific fingerprints Tan et al. (2024). AUNet models dependencies among facial action units and identifies manipulated regions through disruptions of naturally coordinated facial actions Bai et al. (2023).

Overall, existing methods mainly improve generalization by enhancing forgery-sensitive representations or increasing the diversity of training forgeries. While effective, most approaches still focus on appearance-based cues, frequency statistics, or globally aggregated representations. The contextual discrepancy between a facial patch representation and its surrounding facial context remains comparatively underexplored as a direct source of forgery evidence.

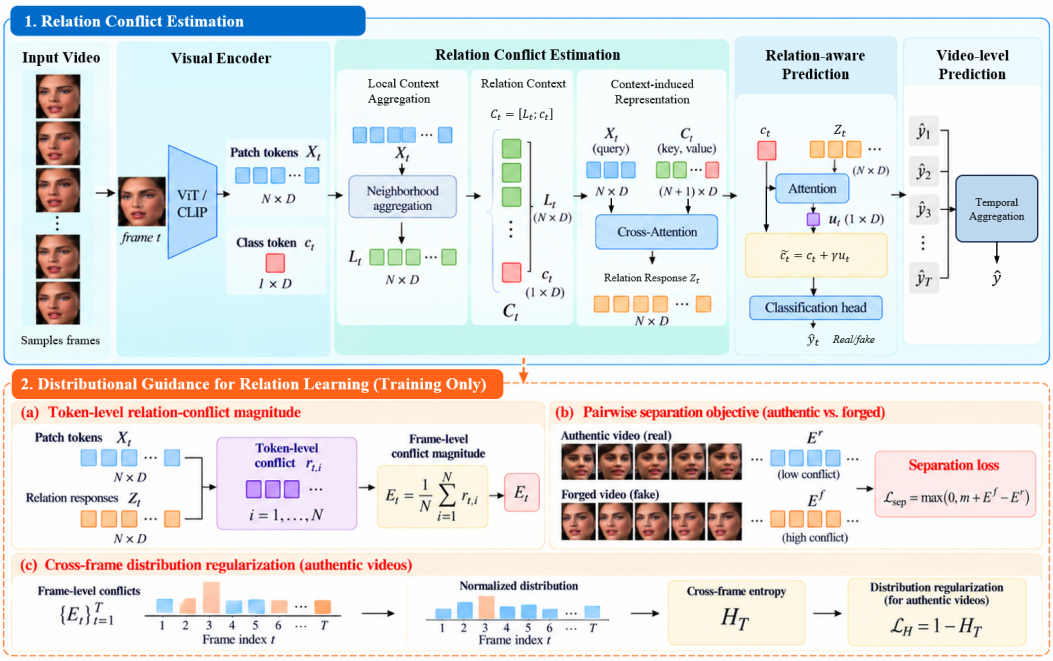


Figure 1: Overview of the proposed Relation Conflict Estimation framework. Context-induced relation responses are constructed from local–global facial context and incorporated into forgery prediction. During training, pairwise conflict separation and cross-frame distribution regularization provide additional guidance for relation learning without extra inference-time computation.

2.2 Contextual and Cross-Frame Consistency Modeling

Video-based forgery detection methods further exploit cross-frame information to capture inconsistencies that are difficult to infer from individual frames. SLF models abnormal variations in style latent flows with a recurrent module Choi et al. (2024), while NACO learns natural spatiotemporal consistency from authentic videos through masked spatial prediction and temporal contrastive learning Zhang et al. (2024). FTCN suppresses spatial shortcuts to emphasize temporal coherence Zheng et al. (2021), whereas AltFreezing alternately optimizes spatial and temporal sub-networks to reduce overfitting to a single cue family Wang et al. (2023d). TALL reorganizes neighboring frames into compact thumbnails, allowing spatial backbones to capture local temporal discontinuities Xu et al. (2023).

More recent studies investigate fine-grained temporal artifacts and contextual aggregation. Video-level blending and spatiotemporal adapter model facial feature drift as transferable temporal evidence Yan et al. (2025b). TFCU decomposes temporal forgery evidence into different forms of accumulated inconsistencies and progressively propagates such evidence across frames Guo et al. (2025b), while FCG combines facial-region-guided spatial learning with temporal modeling to capture subtle forgery traces Han et al. (2025). Beyond explicit temporal modeling, several methods explore global contextual relations and query-based aggregation Cai et al. (2025); Wang et al. (2026a). These methods demonstrate the effectiveness of contextual aggregation, but they primarily focus on enhancing or purifying representations rather than directly characterizing the discrepancy between a facial patch representation and its context-induced response.

3 Method

We formulate face forgery detection as *relation conflict estimation*. The key idea is to characterize the discrepancy between facial patch representations and their context-induced relation responses, and to use such relation evidence for forgery prediction. To improve the generalization of the learned relation representation, we further introduce lightweight distributional guidance on the frame-level

magnitude and cross-frame distribution of relation conflicts. The overall framework is illustrated in Fig. 1.

3.1 Relation Conflict Estimation

Given a face video $V = \{I_t\}_{t=1}^{T}$, let $X_t = \{x_{t,i}\}_{i=1}^{N}$ denote the patch tokens of frame I_t, and c_t denote the corresponding class token. Our motivation is that forgery synthesis may alter the compatibility between a facial patch representation and the contextual information provided by surrounding facial regions. We therefore estimate, for each patch token, a context-induced relation response and define their representation discrepancy as the relation-conflict signal.

To construct the facial context, we first aggregate the 3×3 spatial neighborhood of each patch token to obtain local context tokens L_t, and then combine them with the global class token c_t:

$$C_t = [L_t; c_t]. \tag{1}$$

The local tokens provide neighborhood-level information, while the class token serves as a global semantic reference. We then use the original patch tokens to query the local–global context through cross-attention:

$$Z_t = \mathrm{CrossAttn}(X_t, C_t, C_t), \tag{2}$$

where $Z_t = \{z_{t,i}\}_{i=1}^{N}$ denotes the context-induced relation responses.

The estimated relation responses are directly incorporated into the prediction pathway. Specifically, the class token queries the contextual representation Z_t to aggregate relation evidence:

$$u_t = \mathrm{Attn}(c_t, Z_t, Z_t), \qquad \tilde{c}_t = c_t + \gamma u_t, \tag{3}$$

where γ is a learnable coefficient which controls the contribution of relation evidence. The relation-aware representation $\tilde{c}_t$ is then fed into the classification head for frame-level prediction.

3.2 Distributional Guidance for Relation Learning

The relation responses Z_t provide a contextual representation for each facial patch. We further regularize these representations by explicitly measuring the discrepancy between each patch token and its context-induced response. These regularization terms are applied only during training and incur no additional inference cost beyond the relation-estimation and relation-aware prediction pathway.

Frame-Level Relation-Conflict Regularization. For each patch, we define its relation-conflict magnitude as

$$r_{t,i} = \frac{1 - \cos\left(g(x_{t,i}), g(z_{t,i})\right)}{2}, \tag{4}$$

where $g(\cdot)$ is a lightweight shared projection. A larger $r_{t,i}$ indicates a stronger discrepancy between the original patch representation and its contextual response. We aggregate the token-level conflicts within each frame as

$$E_t = \frac{1}{N} \sum_{i=1}^{N} r_{t,i}, \tag{5}$$

where E_t denotes the frame-level relation-conflict magnitude.

We therefore introduce a separation objective between the two groups without imposing a fixed conflict magnitude on individual forged videos:

$$E^r = \frac{1}{T} \sum_{t=1}^{T} E_t^r, \qquad E^f = \frac{1}{T} \sum_{t=1}^{T} E_t^f, \tag{6}$$

where E^r and E^f denote the video-level relation-conflict magnitudes of the authentic and forged samples, respectively. T is the number of frames.

Since authentic facial regions are expected to remain more compatible with their contextual responses than forged counterparts, we encourage the authentic sample to have a lower relation-conflict magnitude than the forged samples by a margin m. Accordingly, we define the pairwise separation loss as

$$\mathcal{L}_{\mathrm{sep}} = \max\left(0,\, m + E^r - E^f\right). \tag{7}$$

Cross-Frame Distribution Regularization. Besides the average conflict magnitude, we further regularize how frame-level relation conflicts are distributed across the sampled frames. Given the frame-level conflict sequence $\{E_t\}_{t=1}^{T}$, we convert the conflict magnitudes into a normalized cross-frame distribution $\{p_t\}_{t=1}^{T}$ using softmax, where $p_t \propto \exp(E_t)$. We then measure its normalized entropy as

$$H_T = -\frac{1}{\log T} \sum_{t=1}^{T} p_t \log p_t. \tag{8}$$

A higher H_T indicates that relation-conflict responses are more evenly distributed across the sampled frames, whereas a lower value indicates stronger concentration on a subset of frames.

For authentic videos, whose facial-context relations are expected to remain relatively stable across frames, we encourage a less concentrated conflict distribution:

$$\mathcal{L}_H^r = 1 - H_T^r, \tag{9}$$

where H_T^r denotes the normalized cross-frame entropy of an authentic video.

We apply this regularization only to authentic videos. Different forgery mechanisms may produce either sparse or persistent conflict responses across frames. Therefore, imposing a fixed cross-frame distribution on forged videos could introduce manipulation-specific bias. Forged samples are instead jointly constrained through the classification and pairwise relation-conflict separation objectives.

3.3 Overall Objective

During training, all sampled frames are supervised using the video-level label. The classification loss is defined as

$$\mathcal{L}_{ce} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{CE}(\hat{y}_t, y), \tag{10}$$

where $\hat{y}_t$ denotes the prediction of the t-th frame.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{\mathrm{sep}} \mathcal{L}_{\mathrm{sep}} + \lambda_H \mathcal{L}_H^r, \tag{11}$$

where λ_{sep} and λ_H balance the pairwise relation-conflict separation and cross-frame distribution regularization, respectively.

Both auxiliary objectives are derived from the same context-induced relation responses Z_t used for forgery prediction. Specifically, $\mathcal{L}_{\mathrm{sep}}$ encourages authentic videos to exhibit lower relation-conflict magnitudes than their forged counterparts, while $\mathcal{L}_H^r$ regularizes the cross-frame distribution of relation conflicts for authentic videos. The regularization objectives provide training-time distributional guidance without introducing additional inference-time computation. During inference, only the relation estimation and relation-aware prediction pathways are retained, and the frame-level prediction scores are averaged to obtain the final video-level prediction.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Protocols.

We evaluate our method on three face-swapping datasets (Celeb-DF-v2 (CDFv2) Li et al. (2020), DFDC Dolhansky et al. (2020), and DFDCP Dolhansky et al. (2019)), as well as the large-scale DF40 benchmark Yan et al. (2024b), which covers diverse face swapping and AI-generated deepfakes. Following prior work, the model is trained on FaceForensics++ (FF++, c23/HQ) Rossler et al. (2019) and evaluated on remaining datasets. Video-level AUC and frame-level AUC are adopted as the primary evaluation metrics.

Table 1: Cross-dataset evaluation results on the manipulation-based benchmarks. All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	Venue	Video-level AUC			
		CDFv2	DFDC	DFDCP	Avg
AUNet Bai et al. (2023)	CVPR'23	92.77	73.82	86.16	84.25
SLF Choi et al. (2024)	CVPR'24	89.00	-	-	89.00
NACO Zhang et al. (2024)	ECCV'24	89.50	76.70	-	83.10
LSDA Yan et al. (2024a)	CVPR'24	91.10	77.00	-	84.05
MLR Hong et al. (2024)	CVPR'24	91.56	75.17	-	83.36
FPG Xia et al. (2024)	MM'24	94.49	74.75	87.24	85.49
RepDFD Lin et al. (2025)	AAAI'25	89.94	80.99	95.03	88.65
FSFM Wang et al. (2025)	CVPR'25	91.44	83.47	89.71	88.21
TFCU Guo et al. (2025b)	CVPR'25	93.18	86.05	-	89.62
GM-DF Lai et al. (2025)	MM'25	94.80	84.00	**96.30**	91.70
FCG Han et al. (2025)	CVPR'25	95.00	81.80	-	88.40
FFD-StA Yan et al. (2025b)	CVPR'25	94.70	84.30	90.90	89.97
Effort Yan et al. (2025a)	ICML'25	95.60	84.30	90.90	90.27
Forensics Adapter Cui et al. (2025)	CVPR'25	95.70	87.20	92.90	91.93
Ours (ViT-B)	–	95.89±0.46	84.78±0.03	93.88±0.05	91.50±0.15
Ours (CLIP-L)	–	**96.97±0.19**	**88.81±0.12**	96.17±0.02	**93.98±0.11**

Table 2: Cross-manipulation evaluation within the FF++ domain. All methods are trained on FF++ (c23) and tested on unseen face-swapping forgeries. Results are reported in video-level AUC (%).

Method	UniFace	BlendFace	MobileSwap	e4s	FaceDancer	FSGAN	Inswap	SimSwap	Avg
F3Net Qian et al. (2020)	80.90	80.80	86.70	49.40	71.70	84.50	75.70	67.40	74.60
SPSL Liu et al. (2021)	74.70	74.80	88.50	51.40	66.60	81.20	64.30	66.50	71.00
SRM Luo et al. (2021)	74.90	70.40	77.90	70.40	65.90	77.20	79.30	69.40	73.20
CORE Ni et al. (2022)	87.10	84.30	95.90	67.90	77.40	95.80	85.50	72.40	83.30
RECCE Cao et al. (2022)	89.80	83.20	92.50	68.30	84.80	94.90	84.80	76.80	84.40
SLADD Chen et al. (2022)	87.80	88.20	95.40	76.50	82.50	94.30	87.90	79.40	86.50
SBI Shiohara & Yamasaki (2022)	72.40	89.10	95.20	75.00	59.40	80.30	71.20	70.10	76.60
UCF Yan et al. (2023)	83.10	82.70	95.00	73.10	86.20	93.70	80.90	64.70	82.40
IID Huang et al. (2023)	83.90	78.90	88.80	76.60	84.40	92.70	78.90	64.40	81.10
LSDA Yan et al. (2024a)	87.20	87.50	93.00	69.40	72.10	93.90	85.50	79.30	83.50
ProDet Cheng et al. (2024)	90.80	92.90	**97.50**	77.10	74.70	92.80	83.70	84.40	86.70
CDFA Lin et al. (2024)	76.20	75.60	82.30	63.10	80.30	94.20	77.20	75.70	78.10
MAP-Mamba Wang et al. (2026b)	81.50	89.60	95.90	77.90	82.60	-	85.50	81.30	84.90
Effort Yan et al. (2025a)	96.20	87.30	95.30	**98.30**	92.60	95.70	93.60	92.60	94.00
Ours (ViT-B)	93.09	90.07	88.72	82.35	87.42	90.30	89.67	89.36	88.87
Ours (CLIP-L)	**96.84**	**94.59**	97.44	97.37	**98.33**	**97.43**	**97.72**	**96.30**	**97.00**

4.1.2 Implementation Details.

We use ViT-B/16 and CLIP ViT-L/14 as backbones to verify that the proposed method consistently improves performance across backbone architectures of different scales. Notably, for CLIP ViT-L/14, only the visual encoder is used, without involving the text encoder. Input images are resized to 224×224, and $T = 8$ frames are randomly sampled from each video. We set $\lambda_{sep} = \lambda_H^r = 0.05$, the energy margin $m = 0.1$, and the temperature $\tau = 0.5$, empirically. AdamW is used for optimization. All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU. Additional implementation details are provided in the supplementary material.

4.2 Comparison with State-of-the-art Methods

4.2.1 Cross-dataset Evaluation.

We first evaluate our method on manipulation-based benchmarks under the cross-dataset protocol. All models are trained on FF++ and directly evaluated on the unseen CDFv2, DFDC, and DFDCP datasets. As shown in Table 1, our method with CLIP ViT-L/14 achieves AUCs of 96.97%, 88.81%, and 96.17% on CDFv2, DFDC, and DFDCP, respectively, yielding the highest average AUC of 93.98%. Compared with Effort Yan et al. (2025a) and Forensics Adapter Cui et al. (2025), which also adopt CLIP ViT-L as backbone, our method achieves consistent improvements across all three cross-dataset benchmarks, with an average AUC gain of 3.71% and 2.05%. Notably, for CLIP ViT-

Table 3: Cross-generation evaluation results on the generation-based benchmark DF40. All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	VQGAN	StyleGAN2	DDIM	RDDM	DiT	SiT
LSDA Yan et al. (2024a)	71.19	98.69	78.55	77.25	64.68	70.01
RepDFD Lin et al. (2025)	80.50	82.2	-	-	58.7	58.2
Forensics Adapter Cui et al. (2025)	93.90	92.50	-	-	62.0	69.0
MAP-Mamba Wang et al. (2026b)	85.80	92.70	96.70	82.70	-	-
Effort Yan et al. (2025a)	**99.79**	**99.75**	91.49	97.50	88.48	91.56
VLAForge Zhu et al. (2026)	99.70	99.70	-	-	80.30	85.90
Ours(ViT-B)	87.91	98.34	87.28	88.62	67.68	70.42
Ours(CLIP-L)	99.70	99.70	**97.04**	**98.71**	**89.23**	**92.97**

L, only the visual encoder is used, without involving the text encoder, ensuring that the performance gains are attributed to the proposed relation modeling rather than additional textual supervision.

We also demonstrate strong performance when instantiated with ViT-B/16, achieving an average AUC of 91.50%. Despite using a smaller backbone, this variant performs comparably to the CLIP-based Forensics Adapter and outperforms most existing methods. Consistent gains with both ViT-B and CLIP ViT-L/14 demonstrate that relation conflict estimation generalizes across backbone architectures and improves detection of unseen forgeries.

4.2.2 Cross-manipulation Evaluation.

We further evaluate cross-manipulation generalization on the DF40 FS(ff) protocol, where the model is trained on source manipulations and tested on unseen face-swapping methods, including UniFace, BlendFace, MobileSwap, e4S, FaceDancer, FSGAN, InSwap, and SimSwap. This setting is challenging because different face-swapping pipelines introduce heterogeneous artifacts.

As shown in Table 2, our method with ViT-B backbone achieves an average AUC of 88.87% across the eight unseen manipulation pipelines, outperforming representative methods of comparable model scale, including ProDet Cheng et al. (2024), LSDA Yan et al. (2024a), and CDFA Lin et al. (2024). With CLIP ViT-L/14, our method further achieves the highest average AUC of 97.00%, surpassing Effort Yan et al. (2025a), which uses the same CLIP backbone, up to 3%.

These results demonstrate that the proposed method generalizes well across diverse unseen face-swapping manipulations, indicating that the learned relation representations are effective under substantial variation in manipulation pipelines.

4.2.3 Cross-generation Evaluation.

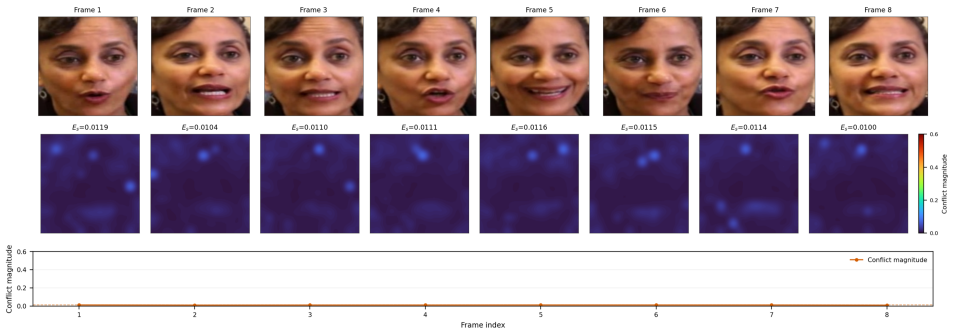
We further evaluate our method on generation-based subsets of DF40, covering GAN-based paradigms (VQGAN and StyleGAN2) and diffusion-based paradigms (DDIM, RDDM, DiT, and SiT). Compared with face swapping, generation-based forgeries often contain fewer explicit blending artifacts and exhibit more realistic appearance, making cross-generation detection particularly challenging.

As shown in Table 3, our method consistently benefits different backbones. With ViT-B, our method shows advantages on GAN-based paradigms, outperforms ViT-B based methods and even achieves competitive performance with the CLIP-based methods Forensics Adapter Cui et al. (2025), especially on StyleGAN2 forgeries. With CLIP ViT-L/14, our method achieves competitive performance with sota on GAN-based paradigms. Importantly, we outperforms existing methods on all four challenging diffusion-based generation paradigms. Compared with Effort Yan et al. (2025a) (CLIP-L backbone), we surpass up to 2% on generation-based subsets, with particularly clear gains on DDIM.

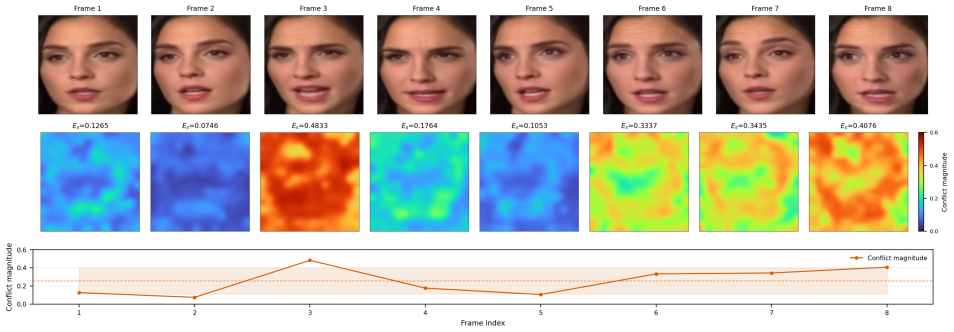
Overall, these results demonstrate that the proposed method generalizes effectively across diverse unseen generative forgeries, including both GAN- and diffusion-based paradigms, highlighting the robustness of the learned relation representations under substantial variation in generation mechanisms.

Table 4: Comparison with vanilla visual backbones in terms of video-level AUC (%).

	ViT-B			CLIP ViT-L/14		
Dataset	Vanilla	Ours	Δ	Vanilla	Ours	Δ
CDFv2	89.30	**95.89**	**+6.59**	90.14	**96.97**	**+6.83**
DFDC	78.14	**84.78**	**+6.64**	82.07	**88.81**	**+6.74**
DFDCP	91.74	**93.88**	**+2.14**	93.19	**96.17**	**+2.98**
Average	86.39	**91.50**	**+5.11**	88.47	**93.98**	**+5.51**



(a) Visualization for an authentic video.



(b) Visualization for a deepfake video.

Figure 2: Visualization of relation-conflict responses for authentic and deepfake videos. For each video, the three rows show sampled frames, conflict heatmaps, and frame-level conflict magnitude, respectively.

4.3 Ablation Studies

4.3.1 Effect of Relation Conflict Representation.

Table 4 compares our method with vanilla ViT-B and CLIP ViT-L/14 backbones trained solely with binary cross-entropy. Our method consistently improves performance across all three datasets, yielding average AUC gains of 5.11% and 5.51% with ViT-B and CLIP ViT-L/14, respectively. Particularly large improvements are observed on CDFv2 and DFDC, while consistent gains are also achieved on DFDCP.

The comparable improvements obtained with both ViT-B and CLIP ViT-L/14 suggest that the gains are not specific to a particular backbone capacity or pretraining scheme. Moreover, the consistent improvements across all three target domains indicate that incorporating contextual relation representations provides complementary discriminative information beyond the original global classification features and effectively improves performances on unseen forgeries.

Table 5: Ablation study of the loss components using the ViT-B backbone. Each variant removes one objective. We report AUC (%) on CDFv2, DFDC, and DFDCP datasets.

Setting	CDFv2	DFDC	DFDCP	Avg.
w/o $\mathcal{L}_{sep}$	91.67	80.86	92.18	88.24
w/o $\mathcal{L}_H^r$	94.29	82.84	92.69	89.94
Full	**95.89**	**84.78**	**93.88**	**91.50**

4.3.2 Effect of Loss Components.

We further analyze the contribution of the proposed training objectives in Table 5. Removing the pairwise separation loss $\mathcal{L}_{\text{sep}}$ causes the average AUC to decrease from 91.50% to 88.24%, with consistent drops on CDFv2, DFDC, and DFDCP, respectively. This indicates that separating the relation-conflict magnitudes of authentic and forged samples provides an important discriminative constraint for relation learning.

Removing the cross-frame distribution regularizer $\mathcal{L}_H^r$ also consistently degrades performance, reducing the average AUC by 1.56%. The decrease is observed across all three target datasets, suggesting that regulating the distribution of relation-conflict responses across sampled frames provides complementary benefits beyond frame-level conflict separation. Overall, the full objective achieves the best performance on all datasets, demonstrating that pairwise conflict separation and cross-frame distribution regularization play complementary roles in improving cross-domain generalization.

4.3.3 Visualization of Relation-Conflict Responses.

Fig. 2 visualizes the learned relation-conflict responses for representative authentic and forged videos. The authentic example exhibits consistently low conflict responses across sampled frames, with a mean conflict magnitude of 0.0111 and a cross-frame standard deviation of only 0.0006. In contrast, the forged example shows substantially stronger and more variable responses, with a mean conflict magnitude of 0.2564 and a standard deviation of 0.1445. This clear separation is consistent with the objective of encouraging lower contextual discrepancy for authentic videos than for forged counterparts.

The response heatmaps further show that authentic frames are dominated by weak conflict activations, whereas forged frames contain broader and stronger high-conflict regions. It can be observed that the relation-conflict responses are higher in the face-swapping and blending regions, which appear as warmer, reddish areas in the heatmaps. This spatial correspondence suggests that the learned relation-conflict responses can highlight regions where facial representations are less compatible with their surrounding context. In addition, he forged example also exhibits noticeably larger frame-to-frame variation, while the authentic sequence remains highly stable. Overall, these visualizations qualitatively support the effectiveness of relation-conflict regularization at both the spatial and cross-frame distribution levels.

5 Conclusion

In this work, we proposed a Relation Conflict Estimation framework for generalizable face forgery detection. The proposed method characterizes contextual discrepancies between facial patch representations and their context-induced relation responses, and integrates the learned relation evidence into forgery prediction. To further improve relation learning, we introduce lightweight distributional guidance that regularizes both relation-conflict magnitude and its cross-frame distribution during training. Extensive experiments demonstrate consistent improvements across different backbones and strong generalization under cross-dataset, cross-manipulation, and cross-generation settings. Ablation studies show that the relation-conflict separation objective provides the major performance gain, while cross-frame distribution regularization offers complementary improvements across target datasets. Qualitative visualizations further reveal that authentic videos exhibit lower and more stable relation-conflict responses, whereas forged videos tend to produce stronger and more variable contextual discrepancies. Overall, these results support relation conflict as useful and transferable evidence for generalizable face forgery detection.

AI USE STATEMENT

In this work, we used generative AI tools for aiding or polishing writing. We have not used generative AI tools for other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24709–24719, 2023.

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 (ICCV)*, pp. 12524–12534, October 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.

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.

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.

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1133–1143, 2024.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217, 2025.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *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:2006.07397*, 2020.

Xinxu Ge, Xin Liu, Zitong Yu, Jingang Shi, Chun Qi, Jie Li, and Heikki Kälviäinen. Difffas: face anti-spoofing via generative diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 144–161, 2024.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–116, 2025a.

Ying Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20818–20827, 2023.

Zonghui Guo, Yingjie Liu, Jie Zhang, Haiyong Zheng, and Shiguang Shan. Face forgery video detection via temporal forgery cue unraveling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7396–7405, 2025b.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22995–23005, 2025.

Cheng-Yao Hong, Yen-Chi Hsu, and Tyng-Luh Liu. Contrastive learning for deepfake classification and localization via multi-label ranking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17627–17637, 2024.

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.

Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8484–8493, 2020.

Yingxin Lai, Hongyang Wang, Jing Yang, Xiangui Kang, Bin Li, Linlin Shen, and Zitong Yu. Gmdf: Generalized multi-scenario deepfake detection. In *ACM International Conference on Multimedia (ACM MM)*, pp. 4300–4309, 2025.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216, 2020.

Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 5262–5270, 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.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021.

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.

Hina Otake, Yoshihiro Fukuhara, Yoshiki Kubotani, and Shigeo Morishima. Detect fake with fake: Leveraging synthetic data-driven representation for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, pp. 316–332, 2024.

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.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 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.

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, 2024.

Changshuo Wang, Jiangming Wang, Ke-Yue Zhang, Taiping Yao, Shouhong Ding, Shunli Wang, Ran Yi, and Lizhuang Ma. Beyond [cls] token: Query-driven token-level forgery purification for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42922–42931, June 2026a.

Chi Wang, Ziwen He, Xinjue Hu, Weinan Guan, Wei Wang, and Zhangjie Fu. Map-mamba: Multi-artifacts perception mamba for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, 21:1184–1197, 2026b.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24364–24376, 2025.

Tianyi Wang and Kam Pui Chow. Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 14548–14556, 2023.

Weihang Wang, Fei Wen, Haoyuan Zheng, Rendong Ying, and Peilin Liu. Conv-mlp: A convolution and mlp mixed model for multimodal face anti-spoofing. *IEEE Transactions on Information Forensics and Security (TIFS)*, 17:2284–2297, 2022a.

Weihang Wang, Peilin Liu, Haoyuan Zheng, Rendong Ying, and Fei Wen. Domain generalization for face anti-spoofing via negative data augmentation. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18:2333–2344, 2023a.

Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7278–7287, 2023b.

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Liang Liu, Yabiao Wang, and Chengjie Wang. Toward high quality facial representation learning. In *31st ACM International Conference on Multimedia (ACM MM)*, pp. 5048–5058, 2023c.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4129–4138, 2023d.

Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4123–4133, 2022b.

Ruiyang Xia, Dawei Zhou, Decheng Liu, Lin Yuan, Shuodi Wang, Jie Li, Nannan Wang, and Xinbo Gao. Advancing generalized deepfake detector with forgery perception guidance. In *ACM International Conference on Multimedia (ACM MM)*, pp. 6676–6685, 2024.

Junhao Xu, Jingjing Chen, Yang Jiao, Jiacheng Zhang, Zhiyu Tan, Hao Li, and Yu-Gang Jiang. Identity-aware vision-language model for explainable face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11278–11286, 2026.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22658–22668, 2023.

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, 2023.

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, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024b.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 70268–70288, 13–19 Jul 2025a.

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, 2025b.

Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision (ECCV)*, pp. 407–424, 2024.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15023–15033, 2021.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15044–15054, 2021.

Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 175–187, 2024.

Jiawen Zhu, Yunqi Miao, Xueyi Zhang, Jiankang Deng, and Guansong Pang. Unleashing vision-language semantics for deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42953–42963, June 2026.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision (ECCV)*, pp. 391–407, 2022.

A Appendix

A.1 Implementation Details

A.1.1 Training Configuration

We optimize all models using AdamW with a weight decay of 0.05. A linear warm-up stage is followed by cosine learning-rate decay. For CLIP ViT-L/14, the base learning rate is set to 2×10^{-5}. The model is trained for 20 epochs with a 5-epoch warm-up. Due to its larger memory footprint, we use a batch size of 1 and accumulate gradients over 8 optimization steps, resulting in an effective batch size of 8. For ViT-B/16, the base learning rate is set to 1×10^{-4}. The model is trained for 100 epochs with a 10-epoch warm-up and a batch size of 12. All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU.

A.2 Additional Results

A.2.1 Complexity Analysis

We further evaluate the computational overhead using a single 224×224 input frame. As shown in Table 6, our method increases 10.05M parameters and 1.8G FLOPs using ViT-B/16 backbone. For CLIP ViT-L/14, the model increases from 304.76M parameters and 81.04G FLOPs to 322.62M and 85.21G, with only 5.86% more parameters and 5.15% more FLOPs. These results show that our method introduces limited computational overhead relative to the backbone during training.

Table 6: Complexity analysis using different visual backbones. FLOPs are measured for a single 224×224 input frame.

Backbone	Setting	Params	FLOPs
ViT-B/16	Baseline	86.69M	17.58G
ViT-B/16	Ours	96.74M	19.38G
CLIP ViT-L/14	Baseline	304.76M	81.04G
CLIP ViT-L/14	Ours	322.62M	85.21G

Table 7: Cross-dataset evaluation results in frame-level AUC (%) for the FFD task. All methods are trained on FF++ and evaluated on other datasets. The best results are indicated in bold, and the second-best results are underlined.

Method	Venue	Frame-level AUC			
		CDFv2	DFDC	DFDCP	Avg
MCF Wang et al. (2023c)	MM'23	73.16	69.63	75.78	72.86
SFDG Wang et al. (2023b)	CVPR'23	75.83	73.64	-	74.74
NoiseDF Wang & Chow (2023)	AAAI'23	75.89	-	63.89	69.89
UCF Yan et al. (2023)	ICCV'23	82.40	-	80.50	81.45
GS Guo et al. (2023)	ICCV'23	84.97	-	81.65	83.31
IID Huang et al. (2023)	CVPR'23	83.80	-	81.23	82.52
LSDA Yan et al. (2024a)	CVPR'24	83.00	73.60	81.50	79.37
RepDFD Lin et al. (2025)	AAAI'25	80.00	77.34	<u>90.57</u>	82.64
FSFM Wang et al. (2025)	CVPR'25	85.05	80.20	85.50	83.58
GM-DF Lai et al. (2025)	MM'25	89.20	<u>84.70</u>	88.20	87.37
TFCU Guo et al. (2025b)	CVPR'25	<u>90.00</u>	84.30	89.00	<u>87.77</u>
Ours	-	**92.26±0.24**	**85.75±0.07**	**93.49±0.05**	**90.50±0.10**

A.2.2 Frame-level Cross-dataset Evaluation

As shown in Table 7, our method achieves the best frame-level cross-dataset performance on all three target datasets, obtaining AUCs of 92.26%, 85.75%, and 93.49% on CDFv2, DFDC, and DFDCP, respectively. Compared with the corresponding second-best results, the improvements are 2.26%, 1.05%, and 2.92% of AUC, leading to an average AUC of 90.50%, which exceeds the previous best average by 2.73%.

Notably, these gains are already observed at the individual-frame level, indicating that the improvement does not rely on video-level score aggregation. The small standard deviations across repeated runs further demonstrate the stability of the proposed method.

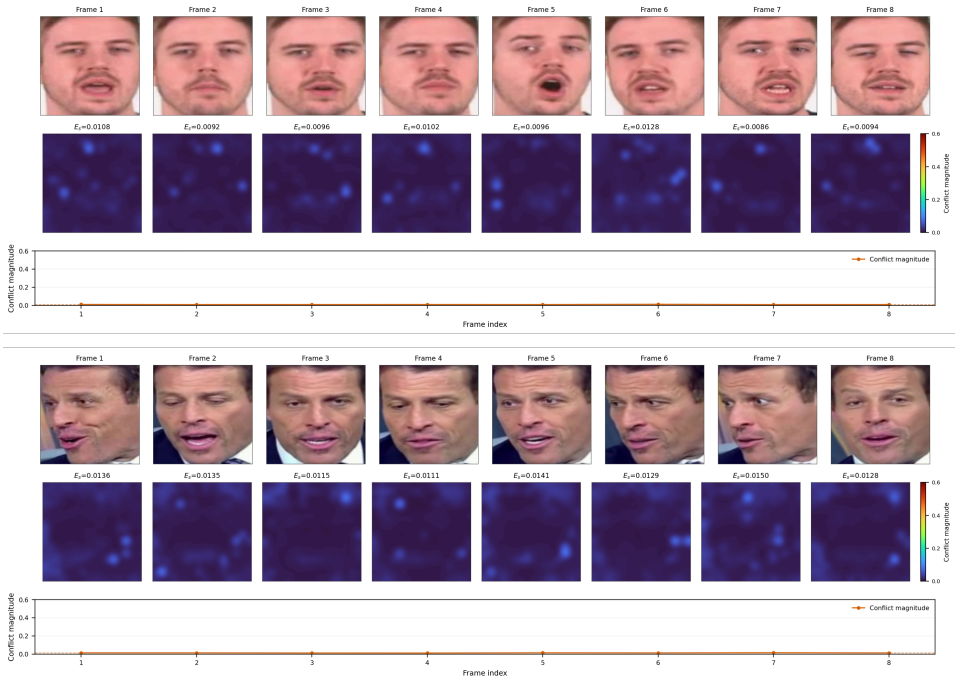
A.2.3 Additional Visualizations

Figure 3 visualizes patch-level relation-conflict responses on authentic and deepfake videos. It can be observed that the relation-conflict responses are higher in the face-swapping and blending regions, which appear as warmer, reddish areas in the heatmaps.

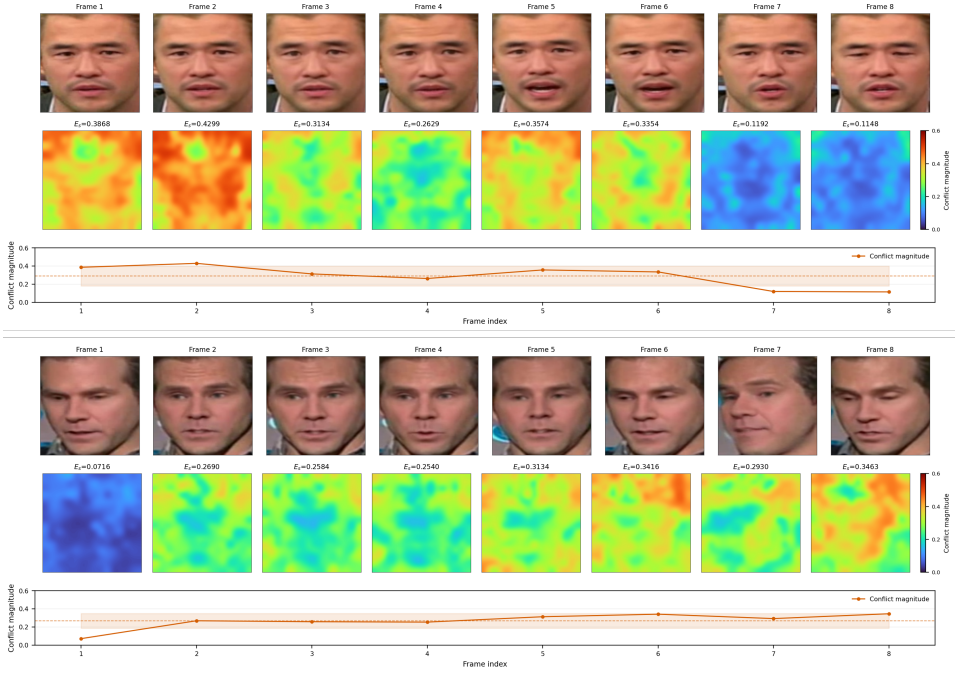
A.3 Dataset Details

A.3.1 Manipulation-based Datasets

FaceForensics++ (FF++). FaceForensics++ Rossler et al. (2019) is a widely used benchmark for face forgery detection. It contains 1,000 pristine videos and manipulated versions generated by four representative facial manipulation methods: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Multiple compression levels are provided to simulate different media qualities. Following the common cross-dataset protocol, we use the c23 (HQ) version of FF++ as the sole training dataset.



(a) Visualization for authentic videos.



(b) Visualization for deepfake videos.

Figure 3: Visualization of relation-conflict responses for authentic and deepfake videos. For each video, the three rows show sampled frames, conflict heatmaps, and frame-level conflict magnitude, respectively.

Celeb-DF-v2 (CDFv2). Celeb-DF-v2 Li et al. (2020) is a large-scale deepfake dataset constructed from celebrity videos collected from YouTube. It employs an improved face-swapping pipeline to reduce obvious synthesis artifacts and produces highly realistic forged videos. Compared with FF++,

CDFv2 contains fewer visible manipulation traces and therefore provides a challenging benchmark for evaluating cross-dataset generalization.

Deepfake Detection Challenge (DFDC). DFDC Dolhansky et al. (2020) is a large-scale and diverse deepfake dataset containing videos with substantial variations in identities, scenes, recording conditions, and manipulation pipelines. Its heterogeneous visual quality and acquisition conditions introduce considerable distribution shifts from FF++, making it particularly suitable for evaluating generalization to real-world forgery scenarios.

DFDC-preview (DFDCP). DFDC-preview Dolhansky et al. (2019) is the preview version of the Deepfake Detection Challenge dataset. It contains both authentic and manipulated videos with diverse subjects, backgrounds, poses, and recording conditions. Although smaller than the full DFDC dataset, its manipulation and acquisition characteristics differ substantially from FF++, providing another representative cross-dataset evaluation benchmark.

A.3.2 Generation-based Datasets

Recent generative models can synthesize entire facial images with high perceptual realism, without relying on the blending and compositing procedures commonly used in conventional face-swapping methods. To evaluate whether our method generalizes beyond manipulation-specific artifacts, we further adopt generation-based subsets from DF40 Yan et al. (2024b). We evaluate six representative generators, including VQGAN, StyleGAN, DDIM, RDDM, DiT, and SiT, which cover substantially different generative architectures and image-generation mechanisms.

VQGAN. VQGAN combines vector-quantized latent representations with adversarial learning for high-quality image synthesis. Its generated faces exhibit globally coherent structures and contain fewer explicit blending artifacts than conventional face-swapping forgeries.

StyleGAN. StyleGAN represents a widely used family of GAN-based generators capable of synthesizing highly realistic and identity-diverse facial images. Its strong perceptual quality makes artifact-based detection increasingly challenging.

DDIM. Denoising Diffusion Implicit Models (DDIM) generate images through an implicit diffusion sampling process. Compared with GAN-based synthesis, diffusion-generated faces exhibit different statistical and frequency characteristics, providing a complementary setting for evaluating cross-generation generalization.

RDDM. RDDM represents another diffusion-based generation paradigm with a distinct reverse-generation process. Its inclusion further increases the diversity of the evaluated diffusion-based forgery distributions.

DiT. Diffusion Transformer (DiT) replaces the conventional convolutional denoising backbone with a Transformer architecture. Its high-quality synthesis and substantially different generative architecture provide a challenging test of whether detectors can generalize beyond artifacts associated with conventional GAN and diffusion backbones.

SiT. SiT is a scalable Transformer-based generative model that formulates image generation through an interpolation-based framework. Together with DiT, it represents recent Transformer-based generative paradigms and enables evaluation on highly realistic synthesized faces with limited conventional manipulation artifacts.