Component-Aware Spatio-Temporal Adaptation of Frozen Foundation Models for Video Deepfake Detection

Tianyi Zhang, Menghan Liang, Gang Zhu, Wenzheng Liu(✉), and Cheng Fu

School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410076, China
liuwenzheng@csust.edu.cn

Abstract. Recent advances in generative video synthesis have made facial manipulations increasingly realistic, which raises growing concerns for media authenticity and digital security. Existing detection methods achieve strong in-domain performance but suffer from cross-dataset degradation, primarily due to overfitting to dataset-specific spatial artifacts. To improve generalization while keeping adaptation efficient, we design an efficient deepfake video detector in which the pretrained foundation encoder remains fixed, while a lightweight decoder is trained to model spatial and temporal forgery cues. First, we introduce a Component-Aware Spatial Enhancement (CASE) module that selectively accentuates manipulation-prone facial regions, such as the eyes, nose and mouth, while capturing global-local structural inconsistencies, thereby enabling the detection of subtle artifacts. Second, it is complemented by a Bidirectional Spatio-Temporal (Bi-ST) decoder that models local temporal transitions and bidirectional temporal dependencies across sampled frames, enabling robust temporal reasoning within each video clip. Without fine-tuning the backbone network, our framework achieves robust cross-dataset generalization by jointly reasoning about spatial and temporal anomalies. Finally, extensive experiments demonstrate that the proposed method performs competitively against strong baselines, particularly under cross-dataset evaluation.

Keywords: Deepfake Detection · Video Forensics · Spatio-temporal modeling · Cross-domain generalization

1 Introduction

Recent progress in deep generative modeling has made realistic facial video synthesis increasingly accessible, giving rise to synthetic videos commonly referred to as deepfakes [8]. While such technology has legitimate applications in creative and entertainment fields, it is increasingly misused for spreading misinformation, identity fraud, and other malicious activities, posing serious threats to social trust, digital security, and media credibility [27].

Early deepfake detection approaches primarily focused on spatial-domain artifacts using convolutional neural networks, with later methods incorporating temporal modelssuch as recurrent networks or Transformers to capture motion inconsistencies [4].

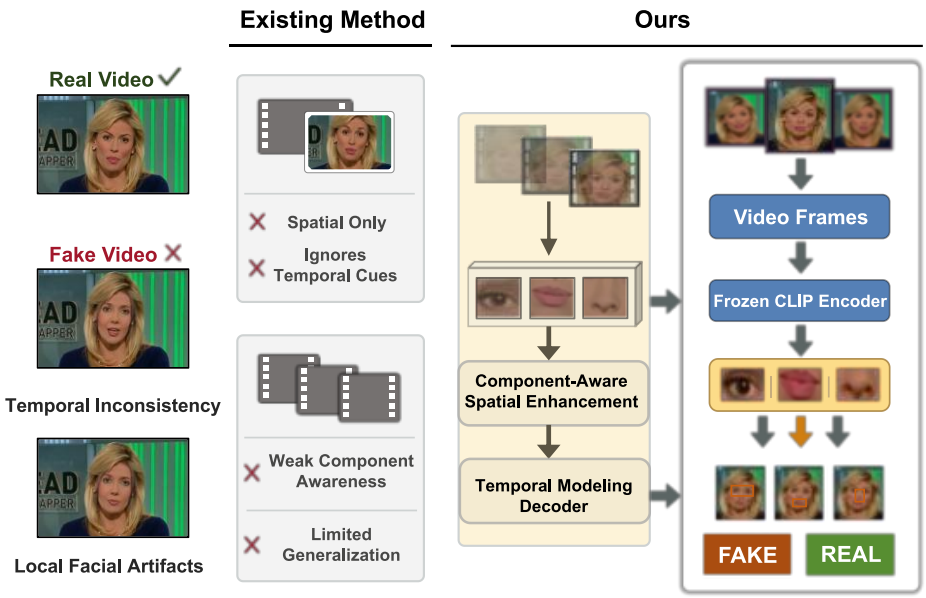

Fig. 1. Motivation and overview of the proposed framework for video-based deepfake detection. Existing methods often overlook component-level artifacts and temporal inconsistencies, while our approach explicitly models both aspects through component-aware spatial enhancement and temporal decoding.

However, good results on the training domain do not necessarily transfer to new datasets. When the test videos come from different manipulation pipelines or acquisition conditions, many detectors become less reliable [16]. As modern deepfake generation techniques effectively suppress low-level artifacts and produce temporally coherent videos, relying on single-modality or dataset-specific cues becomes increasingly insufficient for robust detection.

Recent studies indicate that effective deepfake detection requires jointly modeling manipulation-prone facial regions and temporal inconsistencies while leveraging robust semantic representations [5,17]. However, existing approaches still face notable limitations. Methods focusing on spatial artifacts tend to overfit dataset-specific patterns, while temporal modeling architectures often struggle to capture subtle manipulations and temporal inconsistencies across frames [28]. Moreover, spatial and temporal cues are commonly processed independently, which makes it difficult for the model to learn the interactions among different forgery cues [9]. Although multi-modal fusion strategies have been explored to mitigate these issues [37], they often introduce additional modality dependencies or computational overhead, and may still underutilize complementary cues. These observations suggest that an effective detection framework should prioritize manipulation-sensitive regions, jointly reason over spatial and temporal inconsistencies, and maintain strong generalization without excessive model complexity.

In response to these issues, we design a lightweight video deepfake detector that preserves the pretrained foundation encoder while learning spatio-temporal adaptation modules for forgery recognition. First, we design a Component-Aware Spatial Enhancement module to emphasize semantically meaningful facial regions and capture manipulation-induced global-local discrepancies, enabling the model to focus on the most informative features for forgery detection. This is complemented by a Bidirectional Spatio-Temporal decoder that models temporal inconsistencies across sampled frames through local temporal convolution and bidirectional temporal aggregation, facilitating robust temporal reasoning within each clip. By integrating these components into a unified framework, our approach adapts high-level representations without fine-tuning the backbone, achieving strong cross-dataset generalization. The limitations of existing methods and the advantages of the proposed approach are illustrated in **Fig. 1**. The main contributions are as follows:

- We introduce a Component-Aware Spatial Enhancement module that explicitly focuses on manipulation-prone facial components and captures global-local inconsistencies. By emphasizing semantically meaningful regions, this module improves the detection of subtle artifacts and enhances the interpretability of the model's predictions.
- We design an efficient bidirectional spatio-temporal decoder that aggregates temporal evidence from both forward and backward directions over sampled frames, improving temporal modeling within each clip. This component enhances the modeling of temporal inconsistency patterns and contributes to better performance in challenging cross-dataset scenarios.
- Experimental results on multiple benchmark datasets demonstrate the effectiveness of the proposed spatial and temporal components under complementary evaluation protocols. In particular, the CASE branch achieves strong in-domain performance on FF++, while the full framework delivers robust cross-dataset generalization on several unseen benchmarks.

2 Related Work

2.1 Deepfake Detection from Spatial Artifacts

Most early detectors treated deepfake recognition as an image-level classification problem and relied on CNNs to learn spatial artifacts. MesoNet [1] learns mesoscopic image properties for facial forgery detection, while FaceForensics++ [31] establishes a large-scale benchmark for facial manipulation detection and reports Xception-based detectors as particularly strong baselines. FFD [6] further introduces an attention-based framework for face manipulation detection and localization, showing that local high-frequency fingerprints and manipulated regions are informative for detection. Face X-ray [21] explicitly models blending boundaries through boundary-aware supervision, while SBI [33] improves robustness by training on self-blended images to simulate realistic forgery artifacts.

Beyond purely spatial cues, several works explore frequency-domain representations to capture subtle manipulation traces. F3Net [30] mines frequency-aware clues through frequency-aware decomposition and local frequency statistics, enabling the detection of forgery artifacts that are less visible in the spatial domain. SPSL [24] revisits face forgery detection from a spatial-phase perspective by combining spatial images with phase spectra to capture up-sampling artifacts and improve transferability. SRM [25] further exploits high-frequency noises to improve generalization across manipulation methods. More recently, wavelet-based architectures, such as the wavelet-integrated VGG19 model [32], combine spatial and wavelet-domain features to enhance sensitivity to subtle manipulation traces.

More recently, diffusion-based image generation has posed new challenges to forgery detection. DIRE [39] detects diffusion-generated images by measuring the reconstruction error between an input image and its diffusion-based reconstruction, thereby exploiting intrinsic properties of the diffusion process rather than model-specific artifacts. Nevertheless, these spatial- and frequency-based approaches typically operate at the frame level and may still overfit dataset-specific artifacts, limiting robustness under real-world domain shifts, especially for unseen generative pipelines.

2.2 Video-based Temporal Modeling for Deepfake Detection

To capture inconsistencies beyond single frames, video-based methods explicitly model temporal information across frames. LipForensics [11] exploits temporal irregularities in the mouth region, while FTCN [43] mines inter-frame consistency cues to improve the transferability of video forgery detection. RealForensics [10] leverages self-supervised learning on real talking-face videos to learn natural facial representations without relying on forgery-specific supervision.

Recent approaches adopt more powerful temporal architectures. Methods such as TALL-Swin [41] preserve spatial and temporal dependencies through a thumbnail layout and demonstrate strong cross-dataset performance. Despite these advances, performance degradation under severe domain shifts and unseen manipulation techniques remains a major challenge for video-based detectors.

2.3 Improving Cross-dataset Generalization

Improving cross-dataset generalization has become a central focus in deepfake detection research. CORE [29] proposes consistent representation learning to improve generalization across different manipulation methods. Recce [3] adopts an end-to-end reconstruction-classification framework, where reconstruction learning encourages the model to capture common compact representations of real faces rather than superficial artifacts.

More recent works explore richer representation learning strategies for deepfake detection. Hong et al [13] introduce contrastive learning with multi-label ranking for deepfake classification and localization, Yan et al. [42] propose latent space augmentation to mitigate forgery-specific bias, and Han et al. [12] introduce CLIP-based facial component-guided adaptation for generalizable video-based deepfake detection.

Despite these advances, achieving stable and robust generalization across diverse real-world datasets remains an open problem, motivating further exploration of representation learning beyond dataset-specific cues.

3 Method

We propose a parameter-efficient framework for video-based deepfake detection, which leverages a frozen CLIP image encoder for frame-level spatial representation learning and a lightweight spatio-temporal decoder for temporal reasoning. The framework comprises three key components:

- A CLIP-based multi-layer feature extractor that leverages foundation-level visual representations without backbone fine-tuning.
- A Component-Aware Spatial Enhancement (CASE) module that focuses on manipulation-prone facial components and global–local spatial inconsistencies.
- An efficient Bidirectional Spatio-Temporal (Bi-ST) decoder that models temporal inconsistencies across sampled frames through local temporal convolution and bidirectional temporal aggregation.

The complete pipeline is illustrated in **Fig. 2**, followed by detailed description of its main components in the next subsections.

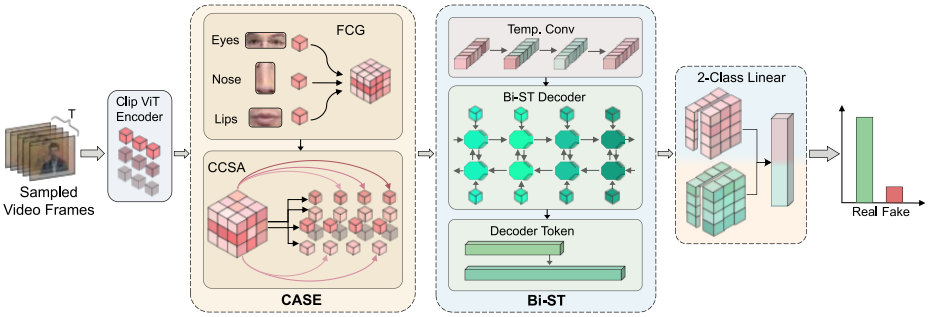

Fig. 2. Overview of the proposed video-based deepfake detection framework. A frozen CLIP image encoder is coupled with a lightweight spatio-temporal decoder. The spatial branch uses CASE, which consists of FCG and CCSA. The temporal branch captures local temporal transitions and broader within-clip dependencies through bidirectional temporal aggregation.

3.1 Problem Formulation

Given an input video $\mathcal{V} = \{I_1, I_2, \ldots, I_T\}$ consisting of T aligned facial frames, where each frame $I_t \in \mathbb{R}^{224\times224\times3}$, the goal is to learn a binary classifier $f: \mathcal{V} \rightarrow \{0,1\}$ that predicts whether the video is real (0) or fake (1). Our approach explicitly models both spatial facial semantics and temporal dependencies across frames.

3.2 CLIP-based Feature Extraction

We employ a frozen CLIP ViT-L/14 image encoder as the visual backbone, leveraging its strong generalization ability from large-scale vision-language pretraining. For a batch of videos, all frames are reshaped into $(B \cdot T, 3{,}224{,}224)$ and independently processed by the encoder.

We extract token embeddings from the last $K = 4$ transformer layers. For each layer l, the output is

$$\boldsymbol{X}^l \in \mathbb{R}^{(B \cdot T) \times N \times D}, \tag{1}$$

where $N = 257$ tokens (one CLS token and 256 patch tokens) and $D = 1024$. The features are reshaped into (B, T, N, D) and forwarded to the decoder. The CLIP encoder remains frozen throughout training to preserve pretrained semantic priors and enable parameter-efficient adaptation.

3.3 Frame-level Detection Pipeline

To clarify the image-frame-level detection process, each aligned RGB face frame $\mathrm{I_t}$ is first encoded by the frozen CLIP image encoder into a global CLS token and local patch tokens. The CLS token captures global facial semantics, while the patch tokens retain local details of facial regions. Based on these tokens, CASE enhances manipulation-prone components, such as the eyes, nose, and mouth/lips, and provides component-aware spatial evidence for forgery detection. In the full framework, the enhanced frame representations are further aggregated by Bi-ST for video-level real/fake classification. For the spatial-only variant reported in **Table 2**, Bi-ST is removed and the prediction is made directly from the CASE-enhanced spatial representation, so as to isolate the contribution of component-aware spatial modeling.

3.4 Component-Aware Spatial Enhancement

Deepfake artifacts are often localized in specific facial regions, while inconsistencies between global facial structure and local details also provide strong cues. To exploit these properties, we introduce the Component-Aware Spatial Enhancement (CASE) module, which integrates Facial Component Guidance (FCG) and Cross-Component Self-Attention (CCSA).

Given the input features $\mathbf{X} \in \mathbb{R}^{B \times T \times N \times D}$, CASE enhances spatial representations through a two-stage process. For clarity, we omit the batch and temporal indices and consider the spatial tokens of a single frame, denoted by $\{\mathbf{x}_j\}_{j=1}^{N-1}$, where $\mathbf{x}_j \in \mathbb{R}^D$. First, component-guided feature weighting is performed using semantic masks $\mathcal{M} = \{M_c : c \in \mathcal{C}\}$ extracted from a pre-trained face parser, where $\mathcal{C} = \{\mathrm{eye}, \mathrm{nose}, \mathrm{mouth}\}$ denotes the set of facial components. For each component $c \in \mathcal{C}$, we compute a scalar importance score and normalize it across spatial tokens:

$$\alpha_{c,j} = \frac{\exp\left(\mathbf{w}_c^\top \mathbf{x}_j + b_c\right)}{\sum_{k=1}^{N-1} \exp\left(\mathbf{w}_c^\top \mathbf{x}_k + b_c\right)} \tag{2}$$

Here, $\mathbf{w}_c$ is the component-specific trainable weight vector, and b_c is the corresponding bias term. The component-guided aggregated feature is then defined as

$$\mathbf{x}_{\mathrm{FCG}} = \sum_{c \in \mathcal{C}} \lambda_c \sum_{j=1}^{N-1} \alpha_{c,j}\, M_c(j)\, \mathbf{x}_j, \tag{3}$$

where $M_c(j) \in \{0,1\}$ indicates whether token j belongs to component c, and $\lambda_c \in \mathbb{R}^+$ is a learnable component importance weight.

Second, cross-component interaction is modeled through bidirectional information exchange between the global representation and local patch representations. Let $\mathbf{g} \in \mathbb{R}^D$ denote the global representation and $\{\mathbf{l}_j \in \mathbb{R}^D : j = 1, \ldots, N-1\}$ denote the local patch representations. The interaction is formulated as

$$\mathbf{g}' = \mathbf{g} + \sum_{j=1}^{N-1} \frac{\kappa(\mathbf{g}, \mathbf{l}_j)}{\sum_{k=1}^{N-1} \kappa(\mathbf{g}, \mathbf{l}_k)} \mathbf{l}_j, \quad \mathbf{l}_j' = \mathbf{l}_j + \frac{\kappa(\mathbf{l}_j, \mathbf{g})}{\sum_{k=1}^{N-1} \kappa(\mathbf{l}_k, \mathbf{g})} \mathbf{g} \tag{4}$$

Here, $\kappa(\cdot,\cdot)$ is defined as

$$\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{\mathbf{x}^\top \mathbf{y}}{\sqrt{d}}\right), \tag{5}$$

where d denotes the feature dimension used in the similarity computation. This kernel form is applied consistently in our implementation, with softmax used for normalization over the relevant comparison set whenever $\kappa(\cdot,\cdot)$ appears in the paper. The enhanced features $\mathbf{X}_{\mathrm{CCSA}}$ are obtained by concatenating the updated global and local representations.

The final CASE output is obtained by integrating the component-guided prior with the cross-component enhanced features, where the latter serves as the main spatial representation for subsequent processing.

3.5 Bidirectional Spatio-Temporal Decoder (Bi-ST)

To capture temporal inconsistencies across sampled frames, we design an efficient Bidirectional Spatio-Temporal decoder, termed Bi-ST. The decoder consists of $K = 4$ stacked layers. Given the input frame features of the i-th Bi-ST layer $\mathbf{X}^{(i)} \in \mathbb{R}^{B \times T \times N \times D}$, where B is the batch size, T is the number of sampled frames, N is the number of tokens per frame, and D is the feature dimension, Bi-ST performs local temporal modeling, bidirectional temporal aggregation, and decoder-token refinement to obtain a compact video-level representation.

Local temporal modeling. To improve the modeling of short-range temporal inconsistencies, we first introduce a lightweight temporal adaptation module over the sampled frame sequence, producing locally enhanced features denoted by $\hat{X}^{(i)}$. In practice, this module is implemented with a depthwise temporal convolution along the temporal

dimension with residual connection, which enhances local frame-to-frame feature transitions while preserving the underlying semantic representation. To explicitly encode frame order, we further add learnable temporal position embeddings:

$$\mathbf{Z}^{(\mathbf{i})}_{b,t,n,:} = \widehat{\mathbf{X}}^{(\mathbf{i})}_{b,t,n,:} + \mathbf{P}^{(\mathbf{i})}_{t,:}, \tag{6}$$

where $\mathbf{P}^{(i)} \in \mathbb{R}^{T\times D}$ is broadcast to all tokens in each frame. In this way, the decoder becomes sensitive to both short-range temporal variation and the temporal position of each frame within the clip.

Bidirectional temporal aggregation. Based on the temporally enhanced features $\boldsymbol{Z}^{(\boldsymbol{i})}$, Bi-ST performs cross-frame interaction in both forward and backward directions. For each frame t, the current representation interacts with its adjacent future frame $(t+1)$ and adjacent past frame $(t-1)$ through a temporal cross-attention module. Rather than introducing explicit scalar coefficients for forward and backward fusion, the bidirectional interactions are jointly modeled within the attention module using direction-specific learnable parameters. The resulting temporally aggregated features are denoted by

$$\mathbf{H}^{(\mathbf{i})}_{t} = \mathbf{Z}^{(\mathbf{i})}_{t} + \mathrm{TCA}\big(\mathbf{Z}^{(i)}_{t}, \mathbf{Z}^{(i)}_{t-1}, \mathbf{Z}^{(i)}_{t+1}\big), \tag{7}$$

where $\mathrm{TCA}(\cdot)$ denotes the bidirectional temporal cross-attention operation, and for boundary frames only the valid temporal direction is used. Compared with one-way temporal propagation, this bidirectional design provides more complete temporal context and facilitates the modeling of subtle inconsistency patterns caused by forgery.

Decoder-token refinement. After bidirectional temporal aggregation, the resulting frame tokens are flattened into a memory sequence and used to update a global decoder token. Specifically, let $\mathbf{x}^{(i)} \in \mathbb{R}^{B\times 1\times D}$ denote the decoder token at the i-th layer. The memory sequence is constructed from all temporally aggregated frame tokens, and the decoder token is refined by attending to this memory:

$$\mathbf{x}^{(i+1)} = \mathrm{Decoder}\left(\mathbf{x}^{(i)}, \mathrm{Flatten}\big(\mathbf{H}^{(i)}\big)\right), \tag{8}$$

where $\mathrm{Flatten}\big(\mathbf{H}^{(i)}\big) \in \mathbb{R}^{B\times(T\times N)\times D}$ denotes the flattened token sequence from all sampled frames, and $\mathrm{Decoder}(\cdot)$ denotes the decoder update block with attention and feed-forward refinement. In this way, temporal evidence from different frames is progressively aggregated into a compact global representation.

After the final Bi-ST layer, the refined decoder token is used as the video-level representation, as it summarizes the temporally aggregated information from the sampled frames. Based on this representation, a two-class linear classifier is applied to produce the final prediction logits for real and fake videos.

4 Experiments

4.1 Implementation Details

In our experiments, CLIP ViT-L/14 is adopted as the visual encoder and its parameters are kept unchanged. Only the proposed CASE branch and Bi-ST decoder are optimized for spatio-temporal adaptation. The model takes $T = 10$ uniformly sampled frames from each 3-second clip, where aligned face crops are resized to 224×224. Following common face-centric video deepfake detection pipelines, facial landmarks are extracted by 2D-FAN, and faces are aligned to the LRW mean face before cropping. During training, we use FF++ (c23) with video-level and frame-level augmentations, including geometric, photometric, and compression perturbations, while validation is conducted on the FF++ validation split without augmentation. The model is trained for 30 epochs using AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-3} on two NVIDIA GeForce RTX 5090 GPUs. Focal loss with $\gamma = 4$ is used for supervision, and early stopping is performed according to the validation AUC.

4.2 Generalization to Unseen Datasets

Table 1. AUC (%) results for cross-dataset evaluation. All models are trained on FF++ and tested on unseen target datasets. Bold and underlined values indicate the best and second-best performance, respectively.

Model	Pub./Mod.	CDF	DFDC	FSh	WDF
Xception [31]	ICCV'19 / Img.	73.7	70.9	72.0	--
Face X-ray [21]	CVPR'20 / Img.	79.5	65.5	92.8	--
LipForensics [11]	CVPR'21 / Vid.	82.4	73.5	97.1	--
FTCN [43]	ICCV'21 / Vid.	86.9	74.0	98.8	--
RealForensics [10]	CVPR'22 / Vid.	86.9	75.9	<u>99.3</u>	--
AltFreezing [40]	CVPR'23 / Vid.	89.5	--	<u>99.3</u>	--
TALL-Swin [41]	ICCV'23 / Vid.	90.8	76.8	**99.7**	--
Yan et al. [42]	CVPR'24 / Img.	91.1	<u>77.0</u>	--	--
Choi et al. [4]	CVPR'24 / Vid.	89.0	--	99.0	--
Hong et al. [13]	CVPR'24 / Img.	91.6	75.2	--	73.4
Nguyen et al. [28]	ICCV'25 / Vid.	<u>92.4</u>	74.6	--	74.2
Kim et al. [17]	ICCV'25 / Vid.	89.7	75.2	<u>99.3</u>	--
MoE-FFD [18]	TDSC'25 / Vid.	91.3	--	--	<u>83.9</u>
Ours	--	**93.6**	**82.6**	97.7	**84.5**

We examine the cross-dataset robustness of the proposed method by evaluating it on multiple unseen deepfake benchmarks without using any target-domain samples during training.

Evaluation Benchmarks. Following the cross-dataset protocol in RealForensics [10], we evaluate our method on multiple representative deepfake benchmarks. FaceForensics++ (FF++) [31] is used for training, while CelebDF-v2 (CDF) [22], DFDC [7], FaceShifter (FSh) [20], and WildDeepfake (WDF) [44] are used as unseen target datasets. For all cross-dataset results, we report performance only on the official test/evaluation split of each target dataset, without using any target-domain samples during training or model selection. These benchmarks cover both controlled and in-the-wild scenarios, providing a rigorous testbed for evaluating cross-domain generalization.

Cross-dataset Evaluation. As shown in **Table 1**, our method demonstrates strong and balanced generalization across unseen datasets. It achieves the best AUC on **CDF (93.6%), DFDC (82.6%),** and **WDF (84.5%)**. Specifically, our method outperforms Nguyen et al. [28] on CDF by **1.2%**, surpasses Yan et al. [42] on DFDC by **5.6%**, and improves over MoE-FFD [18] on WDF by **0.6%**. Although TALL-Swin [41] achieves the highest AUC on **FSh**, our method still obtains a competitive result of **97.7%**. Overall, these results suggest that our framework provides more consistent cross-dataset generalization across benchmarks with diverse characteristics, including both controlled and in-the-wild scenarios.

4.3 Per-manipulation Results on FF++

Table 2. Intra-testing results trained on FF++, measured by AUC. The best and second-best results are shown in bold and underlined, respectively.

Model	Pub.	DF	F2F	FS	NT	AVG.
EfficientB4 [36]	PMLR'19	97.6	97.6	98.0	93.1	96.6
CNN-Aug [38]	CVPR'20	90.5	87.9	90.3	73.1	85.4
F3Net [30]	ECCV'20	97.9	98.0	98.4	93.5	97.0
FFD [6]	CVPR'20	98.0	97.8	98.5	93.1	96.9
SPSL [24]	CVPR'21	97.8	97.5	98.3	93.0	96.7
SRM [25]	CVPR'21	97.3	97.0	97.4	93.0	96.2
CORE [29]	CVPR'22	97.9	98.0	98.2	93.4	96.9
Recce [3]	CVPR'22	98.0	97.8	97.9	93.6	96.8
ADA-FInfer [14]	TDSC'24	83.1	78.4	74.4	72.3	77.1
VGG19 [32]	BigData'24	97.0	96.4	97.7	91.4	95.6
SDIF [19]	ICASSP'24	83.5	83.7	75.1	--	--
DDL [35]	TIFS'25	83.1	77.5	78.1	79.3	79.5
CAFMF [26]	ICCRD'25	<u>98.8</u>	98.2	98.9	<u>94.4</u>	<u>97.6</u>
FRENet [34]	IVC'25	98.6	<u>98.3</u>	<u>99.0</u>	93.8	97.4
Ours (w/o Bi-ST)	--	**99.3**	**99.3**	**99.0**	**99.1**	**99.2**

We also evaluate the proposed method on different manipulation types in FaceForensics++, including DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). The intra-testing AUC results are summarized in **Table 2**. Since many recent methods have already approached saturation on the in-domain FF++ benchmark, introducing additional video-level temporal modeling in this setting may make it difficult to determine whether the improvement comes from stronger frame-level forgery representation or simply from extra temporal aggregation. Therefore, we use **Ours (w/o Bi-ST)** for comparison with the frame-based detectors in **Table 2**, such as EfficientB4, F3Net, CORE, Recce, CAFMF, and FRENet, so as to maintain a comparable frame-level evaluation protocol. This allows us to more clearly isolate the contribution of the proposed component-aware spatial modeling without the influence of additional temporal modeling. Even under this more controlled comparison, **Ours (w/o Bi-ST)** achieves the best overall performance, obtaining **99.3%**, **99.3%**, **99.0%**, and **99.1%** AUC on DF, F2F, FS, and NT, respectively, with an average AUC of **99.2%**. Compared with the strongest recent frame-based baselines, our method improves the average AUC over CAFMF and FRENet by **1.6%** and **1.8%**, respectively. The improvement on NT is particularly notable, where our method surpasses CAFMF and FRENet by **4.7%** and **5.3%**, respectively. These results indicate that the proposed spatial branch alone is already capable of capturing highly discriminative and manipulation-agnostic forgery cues. In other words, the superiority of our framework does not rely solely on temporal modeling; rather, its frame-level spatial representation is already strong enough to outperform existing frame-based detectors, while the Bi-ST module is mainly beneficial for more challenging generalization scenarios.

4.4 Parameter Efficiency

Table 3. Comparison of trainable parameters (#T.P.) and cross-dataset AUC (%) performance under the CLIP ViT-L/14 backbone.

Method	#T.P.	CDF	DFDC	FSh	WDF	Avg.
LCP [2]	2.0K	75.8	77.0	88.0	82.1	80.7
VPTShallow [15]	3.1K	78.9	76.4	89.0	82.9	81.8
VPTDeep [15]	26.6K	81.1	<u>77.5</u>	<u>90.0</u>	<u>86.7</u>	<u>83.8</u>
EVL [23]	58.3M	80.3	75.8	87.7	79.1	80.7
FFT	303.0M	<u>88.3</u>	75.2	84.8	83.8	83.0
Ours	**1.69M**	**93.6**	**82.6**	**97.7**	**84.5**	**89.6**

To further examine the parameter efficiency of our framework, we compare it with several CLIP ViT-L/14-based adaptation baselines, including linear classifier probing, visual prompt tuning, efficient video learning, and full fine-tuning. The corresponding results are reported in **Table 3**. For VPT, both the shallow and deep variants are considered. Unlike LCP and VPT, which mainly rely on the final representation of a frozen CLIP encoder, our method exploits intermediate-layer attributes to build a structured

spatial-temporal side network, where facial component priors guide spatial modeling and cross-frame inconsistency is captured for temporal modeling. Different from EVL, our framework further introduces a face-specific and more interpretable spatial enhancement mechanism rather than relying primarily on generic feature aggregation. Therefore, while using more trainable parameters than LCP and VPT, our method remains substantially lighter than EVL and full fine-tuning, achieving a better balance between efficiency and cross-dataset generalization.

According to the results in **Table 3**, our approach outperforms the compared adaptation strategies on CDF, DFDC, FSh, and overall average performance, with only 1.69M trainable parameters. These results indicate that our method preserves the advantage of parameter-efficient adaptation while achieving superior cross-dataset generalization.

4.5 Ablation Study

We analyze the contributions of the Component-Aware Spatial Enhancement (CASE) module and the Bidirectional Spatio-Temporal (Bi-ST) decoder under cross-dataset evaluation in **Table 4**.

Table 4. Ablation study of the proposed framework in terms of AUC (%). CASE: Component-Aware Spatial Enhancement module. Bi-ST: Bidirectional Spatio-Temporal decoder.

Comp.	CDF	DFDC	FSh	WDF	Avg.
Ours	93.6	82.6	97.7	84.5	89.6
w/o CASE	89.9	83.2	97.9	82.1	88.3(-1.3)
w/o Bi-ST	93.3	79.3	96.4	80.3	87.3(-2.3)

Effect of CASE. Removing CASE causes an average AUC drop of **1.3%**, with clear degradation on CDF and WDF. This suggests that CASE helps the model focus on semantically meaningful and manipulation-prone facial regions, thereby improving spatial discriminability and reducing irrelevant background bias. As further illustrated in **Fig. 3**, the model with CASE produces more concentrated and discriminative affinity responses on key facial regions, whereas the variant without CASE exhibits more diffuse attention.

Effect of Bi-ST. Removing Bi-ST leads to a larger average AUC drop of **2.3%**, with the most obvious decline on DFDC and WDF. This verifies the importance of temporal modeling for capturing cross-frame inconsistency patterns that cannot be reliably inferred from spatial cues alone.

Discussion on dataset-specific fluctuations. We also note that removing one module yields slightly better results on a few individual datasets. We consider these variations to be dataset-specific fluctuations caused by distribution differences across unseen

benchmarks, rather than evidence against the usefulness of the removed component. Different datasets may rely on different dominant forgery cues, so improvements brought by a specific module do not necessarily appear uniformly on every benchmark. More importantly, the full model achieves the best overall average AUC, demonstrating that CASE and Bi-ST provide complementary benefits: CASE strengthens structured spatial perception, while Bi-ST enhances temporal inconsistency modeling. Their organic combination leads to a more balanced and transferable representation, which is crucial for robust cross-dataset generalization.

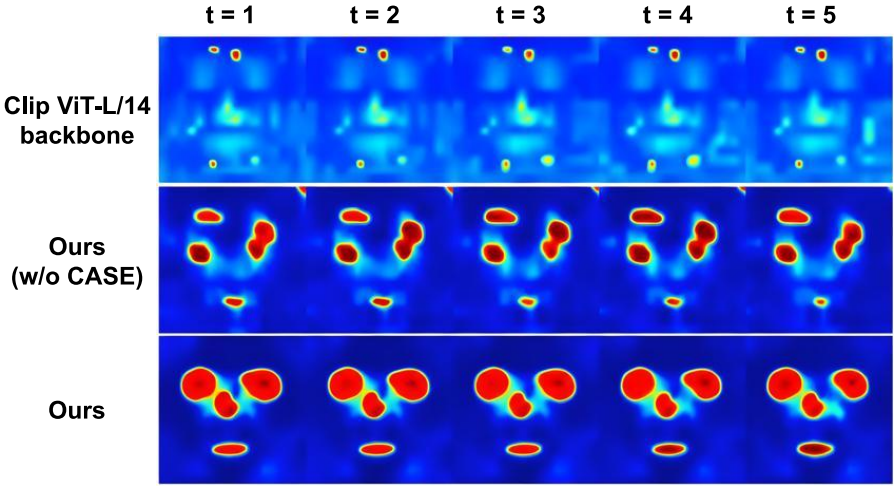

Fig. 3. Frame-wise affinity map comparison among three variants: the baseline model, our model without CASE, and the full model equipped with CASE.

5 Conclusion

This paper presents a parameter-efficient, video-based deepfake detection framework designed to address the critical challenges of cross-dataset generalization and robust spatio-temporal modeling. By utilizing a frozen foundation model encoder, our approach retains powerful, general-purpose semantic representations without incurring the computational cost or overfitting risks of full fine-tuning. The proposed CASE module directs the model's attention to semantically meaningful and manipulation-prone facial regions, effectively capturing subtle global-local structural discrepancies. Complementing this, the Bi-ST decoder models local temporal transitions and broader within-clip temporal inconsistencies across sampled frames, enabling more reliable temporal reasoning within each sampled clip. Extensive evaluation on multiple cross-dataset benchmarks demonstrates that our framework achieves competitive performance and strong robustness under distribution shifts. For future work, we plan to explore more adaptive mechanisms for facial component discovery, investigate richer multi-modal fusion strategies, develop efficient temporal modeling for longer

sequences, and adapt the framework for real-world deployment scenarios involving streaming data and continuously evolving manipulation techniques.

Acknowledgments. This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (Grant No. 24B0308), the National Innovation Training Program for College Students, China (Grant No. 202510536030).

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS). pp. 1–7. IEEE (2018)
2. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
3. Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: 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)
4. Choi, J., Kim, T., Jeong, Y., Baek, S., Choi, J.: Exploiting style latent flows for generalizing deepfake video detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1133–1143 (2024)
5. Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J.: Forensics adapter: Adapting clip for generalizable face forgery detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19207–19217 (2025)
6. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition. pp. 5781–5790 (2020)
7. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
8. Gambín, Á.F., Yazidi, A., Vasilakos, A., Haugerud, H., Djenouri, Y.: Deepfakes: current and future trends. Artificial Intelligence Review 57(3), 64 (2024)
9. Guo, Z., Liu, Y., Zhang, J., Zheng, H., Shan, S.: Face forgery video detection via temporal forgery cue unraveling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7396–7405 (2025)
10. Haliassos, A., Mira, R., Petridis, S., Pantic, M.: Leveraging real talking faces via self-supervision for robust forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14950–14962 (2022)
11. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5039–5049 (2021)
12. Han, Y.H., Huang, T.M., Hua, K.L., Chen, J.C.: Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22995–23005 (2025)
13. Hong, C.Y., Hsu, Y.C., Liu, T.L.: Contrastive learning for deepfake classification and localization via multi-label ranking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17627–17637 (2024)

14. Hu, J., Liang, J., Qin, Z., Liao, X., Zhou, W., Lin, X.: Ada-finfer: Inferring face representations from adaptive select frames for high-visual-quality deepfake detection. IEEE Transactions on Dependable and Secure Computing (2024)
15. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European conference on computer vision. pp. 709–727. Springer (2022)
16. Kashiani, H., Alipour, N., Afghah, F.: Freqdebias: towards generalizable deepfake detection via consistency-driven frequency debiasing (2025)
17. Kim, T., Choi, J., Jeong, Y., Noh, H., Yoo, J., Baek, S., Choi, J.: Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11198–11207 (2025)
18. Kong, C., Luo, A., Bao, P., Yu, Y., Li, H., Zheng, Z., Wang, S., Kot, A.C.: Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. IEEE Transactions on Dependable and Secure Computing (2025)
19. Lai, Y., Yang, G., He, Y., Luo, Z., Li, S.: Selective domain-invariant feature for generalizable deepfake detection. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2335–2339. IEEE (2024)
20. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5074–5083 (2020)
21. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5001–5010 (2020)
22. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3207–3216 (2020)
23. Lin, Z., Geng, S., Zhang, R., Gao, P., De Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: European Conference on Computer Vision. pp. 388–404. Springer (2022)
24. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: 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)
25. Luo, Y., Zhang, Y., Yan, J., Liu, W.: 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)
26. Meng, J., Dong, J., Liu, Y., Liang, C.: The application of cross-modal consistency enhancement and feature fusion in high-frequency feature-driven face forgery detection. In: 2025 IEEE 17th International Conference on Computer Research and Development (ICCRD). pp. 111–116. IEEE (2025)
27. Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., Dwivedi, Y.K.: Deepfakes: Deceptions, mitigations, and opportunities. Journal of Business Research 154, 113368 (2023)
28. Nguyen, D., Astrid, M., Kacem, A., Ghorbel, E., Aouada, D.: Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2025)
29. Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., Zhao, Y.: 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)
30. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European conference on computer vision. pp. 86–103. Springer (2020)

31. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Face forensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
32. Sadhya, S., Qi, X.: Deepfake detection with wavelet-integrated convolutional networks. In: 2024 IEEE International Conference on Big Data (BigData). pp. 4540–4547. IEEE (2024)
33. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18720–18729 (2022)
34. Song, W., Guo, S., Gao, M., Li, Q., Zhu, X., Rida, I.: Deepfake detection via feature refinement and enhancement network. Image and Vision Computing p. 105663 (2025)
35. Sun, Z., Ruan, N., Li, J.: Ddl: Effective and comprehensible interpretation framework for diverse deepfake detectors. IEEE Transactions on Information Forensics and Security (2025)
36. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
37. Wang, L., Zhao, J., Zhang, X., Guo, X., Yuan, Y., Zhang, T., Chu, J., Jiang, Y., Yang, X., Jin, L., et al.: Erf-ba-tfd+: a multimodal model for audio-visual deepfake detection. Vicinagearth 2(1), 1–12 (2025)
38. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
39. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22445–22455 (2023)
40. Wang, Z., Bao, J., Zhou, W., Wang, W., Li, H.: Altfreezing for more general video face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4129–4138 (2023)
41. Xu, Y., Liang, J., Jia, G., Yang, Z., Zhang, Y., He, R.: Tall: Thumbnail layout for deepfake video detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22658–22668 (2023)
42. Yan, Z., Luo, Y., Lyu, S., Liu, Q., Wu, B.: Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8984–8994 (2024)
43. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: 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)
44. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2382–2390 (2020)