Towards More General Video-based Deepfake Detection through Facial Component Guided Adaptation for Foundation Model

Supplementary Material

8. More Experiments for Model Analysis

In this section, we provide additional experiments to further analyze our framework. We remain the Area Under the receiver operating characteristic Curve (AUC) as the evaluation metric in the experiments.

8.1. Importance of Facial Components in the FCG

In Table 7, we evaluate the cross-dataset performance of the models by excluding specific facial components from our Face Component Guidance (FCG) mechanism. We observe that excluding guidance for any component results in a slight decrease in average performance. Notably, the greatest performance drop occurs when the 'eyes' facial component is excluded, suggesting it more critical than other facial components for generalization. Consequently, we include all facial components in our FCG, as this approach achieves the best overall performance across all datasets.

8.2. The Selection of the Target Attribute

In Table 8, we explore different target attribute γ^s settings to evaluate the effectiveness of attributes to improve model generalizability to focus on major facial parts. We can see in the table that selecting k and q both performs well in improving generalizability while v have a degraded performance. We surmise this due to the design nature of q and k in for affinity evaluations. In our experiments, we set $\gamma^s = k$ due to its slightly advanced performance.

8.3. Evaluation on Videos of AI Avatars

To further validate our model's generalizability to modern AI video generators, we collect a dataset comprising 56 videos from 28 different content creators who use **HeyGen**[5] to generate videos with their AI avatars. We evaluate these collected videos using our method, RealForensics [13], and LAA-Net (w/ SBI) [33] for comparison. All models are pre-trained on the FF++ dataset, and the scores from previous works are evaluated using their official code and model checkpoints under the default settings. Our method achieves an AUC of 86.1%, while the RealForensics method achieves an AUC of 84.9%, and the LAA-Net (w/ SBI) method achieves only 45.4% AUC. The superior performance of our approach compared to previous SOTA methods further demonstrates its enhanced generalizability.

[5]https://www.heygen.com/

Table 7. **Importance of Facial Components in FCG:** We evaluate the cross-dataset performance on models trained with excluded facial components in the FCG. This experiment demonstrates the impact of each facial parts to improve model generalization.

Method	CDF	DFDC	FSh	DFo	WDF	Avg.
Ours	**95.0**	**81.8**	**98.1**	**99.6**	**87.2**	**92.3**
w/o eyes	93.9	81.2	98.1	99.3	85.8	91.5
w/o nose	94.3	81.7	97.1	99.3	86.0	91.7
w/o lips	94.5	81.5	97.8	**99.6**	86.7	92.0
w/o skin	94.1	81.6	97.4	99.5	87.1	91.9

Table 8. **Evaluation of γ^s Parameter**: We select different γ^s values in $\{q, k, v\}$ to evaluate the efficacy for attributes to collect informative facial features.

γ^s	CDF	DFDC	FSh	DFo	WDF	Avg.
$\gamma^s = k$	95.0	**81.8**	**98.1**	**99.6**	**87.2**	**92.3**
$\gamma^s = q$	**95.1**	81.5	97.8	99.2	86.7	92.1
$\gamma^s = v$	84.2	81.6	97.9	99.5	86.0	89.8

9. Evaluation on Modern Deepfake Techniques

Beyond comprehensive comparisons on video-based Deepfake detection, we also evaluate the proposed approach on unseen novel Deepfakes, with a particular focus on recent Diffusion models. Since the latest advancements in Deepfakes using Diffusion models are image-based, we adapt our video-based framework to operate under similar conditions by removing the temporal module and retaining only the spatial module in the decoder block.

To ensure fair comparison, we followed the protocol from DIRE [50] and utilize the CelebA-HQ subset from their proposed DiffusionForensics [50] dataset, which contains facial images generated by Diffusion Models (DMs). It includes real images from CelebA-HQ [19] and fake images generated by SD-v2 [40] as the training subset, while the testing subset further includes images generated from IF [43], DALLE-2 [39], and Midjourney. The results of the following experiments are reported using the Average Precision (AP) metric, expressed as a percentage.

Generalizability to Diffusion Models. In Table 9, we evaluate the effectiveness of our framework on images generated by novel diffusion models (DMs). In the upper section, we assess the zero-shot capability of our framework alongside the State-Of-The-Art (SOTA) image-based SBI method. Both methods are pre-trained on the FF++ dataset and evaluated on the testing subset from DiffusionForen-

Table 9. **Evaluation on Novel Diffusion Deepfakes**: In the upper table, we evaluate the zero-shot capability of our framework with the SBI. In the bottom table, we compare with methods trained on the CelebA-HQ split of the DiffusionForensics dataset.

Method	Generated face images				
	SD-v2	IF	DALLE-2	Midjourney	Avg.
SBI [44]	70.8	83.9	64.4	41.5	65.2
Ours	**96.8**	**93.1**	**71.4**	**62.7**	**81.0**
CNNDet [49]	99.8	82.7	33.7	69.3	71.4
F3Net [37]	99.1	84.9	69.8	87.9	85.4
DIRE [50]	**100**	99.9	**99.9**	**100**	**100**
Ours	**100**	**100**	99.8	**100**	**100**

sics. The results demonstrate that our framework exhibits stronger zero-shot capability, outperforming the SOTA SBI by an average of 15.8% AP. This performance can be attributed to the FCG, which prevents the model from overfitting to dataset-specific cues.

In the lower section, we follow the protocols of previous methods to train and evaluate our framework on the DiffusionForensics dataset. We compare our approach against prior methods (CNNDet [49], F3Net [37], and DIRE [50]) to demonstrate its generalizability. The results in the table show that our framework achieves performance on par with the SOTA method (DIRE).

Robustness Towards Common Perturbations. In real-world scenarios, images often undergo various post-processing adjustments, making robustness to unseen perturbations crucial. In this section, we evaluate the robustness of our framework against two types of disturbances: Gaussian blur ($\sigma = 0, 1, 2, 3$) and JPEG compression (quality = 100, 65, 30). We follow the evaluation setup from the previous section to assess robustness under both zero-shot and in-domain regimes. The results are presented in Table 10. Our model demonstrates strong robustness, with no significant performance degradation under these perturbations, particularly in the zero-shot evaluation.

10. Elaboration on Attribute Extraction

To elucidate the connection between attention attributes and patch embeddings within the Vision Transformer (ViT) encoder pipeline, we detail the workflow of a typical ViT encoder layer in Algorithm 1. Each layer accepts embeddings from the preceding layer as input, which encompasses a class embedding along with numerous patch embeddings. These embeddings are then processed through a self-attention mechanism to produce output embeddings for the subsequent layer. Initially, the class embedding is represented by a learnable token, and patch embeddings are formed by a distinct patch extraction layer given an image (for further details, please see the ViT paper [10]). In the algorithm presented, W_s and B_s, for $s \in \{Q, K, V, O\}$, signify the weights and biases associated with the linear transformations. LN_1 and LN_2 represent the layer normalization modules. The ***MHSA*** stands for the Multi-Head Self-Attention mechanism, which operates on the query, key, and value embeddings of the class and patch tokens. Furthermore, the MLP (multi-layer perceptron) includes two linear layers and a GeLU activation layer. The attention attributes $\mathcal{A}_{l,\gamma}$, where $\gamma \in \{q, k, v\}$, are extracted as specified in the cited lines 4, 5, and 6, and the extracted patch embeddings $\mathcal{P}_l$ are referred to in line 10.

Table 10. **Robustness on Novel Diffusion Deepfakes:** We assess the zero-shot and in-domain robustness of our framework with SBI and DIRE, respectively.

Method	JPEG (Quality)			Blur (Sigma)				Avg.
	100	65	30	0	1	2	3	
SBI	65.2	53.5	56.3	65.2	51.3	52.9	54.2	55.6
Ours	**100**	**79.3**	**76.2**	**100**	**81.5**	**81.1**	**79.2**	**85.3**
DIRE*	**100**	99.8	99.8	**100**	**99.9**	**99.9**	**99.9**	**99.9**
Ours*	**100**	**100**	**99.9**	**100**	**99.9**	**99.9**	**99.9**	**99.9**

Algorithm 1 The workflow of the ViT transformer encoder layer given embeddings from the previous layer.

1: **Input** x
2: **Output** emb
3: $\hat{x} \leftarrow LN_1(x)$
4: $q \leftarrow W_Q\hat{x} + B_Q\hat{x}$ ▷ Query Transform
5: $k \leftarrow W_K\hat{x} + B_K\hat{x}$ ▷ Key Transform
6: $v \leftarrow W_V\hat{x} + B_V\hat{x}$ ▷ Value Transform
7: $z \leftarrow \boldsymbol{MHSA}(q, k, v)$ ▷ Multi-head Self-Attention
8: $out \leftarrow W_O z + B_O z$
9: $x' \leftarrow x + out$
10: $emb \leftarrow x' + MLP(LN_2(x'))$

11. Inference Time

The average inference time of our framework on a 3-second video is 1.5 seconds using an A5000 GPU. As our framework leverages the CLIP image encoder to extract generic features for adaptation, most of the inference time is spent in the image encoder processing the 10 frames extracted from the video clip. In contrast, our proposed lightweight decoder modules require minimal inference time.

12. Societal Impact Concern

Since the proposed method mainly works on Deepfake detection problem to mitigate the negative influences brought by Deepfake technologies, there is no major societal impact concerns.

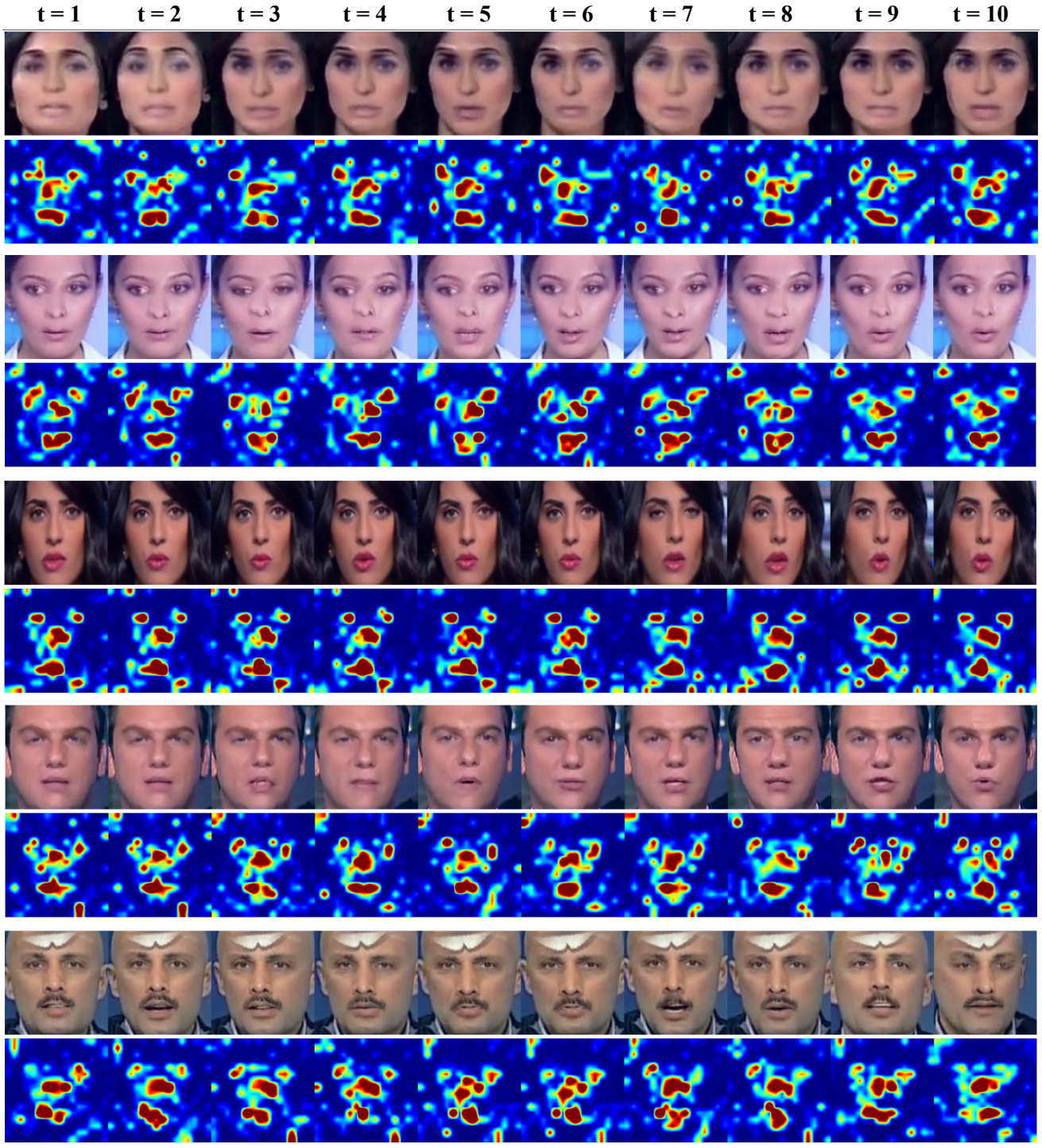

Figure 5. **Attention Visualization for Individuals:** We present the input frames along with the per-frame attention affinity map for individual subjects. We retain the experimental settings described in Sec. 4.8 while sampling only a single clip for visualization.