Leave No Stone Unturned: Uncovering Holistic Audio-Visual Intrinsic Coherence for Deepfake Detection

Supplementary Material

A. Overview

In this supplementary material, we provide additional details of the proposed HAVIC framework and the HiFi-AVDF dataset, as well as more experimental results. Sec. B introduces further implementation details, including model architecture choices and training configurations. Sec. C then offers a comprehensive description of the proposed HiFi-AVDF dataset and information about other datasets. Finally, Sec. D reports additional results, including extended evaluations, ablation studies, and qualitative analysis.

B. Implementation Details

B.1. Holistic Coherence Priors Pre-training

Inputs. We sample 3.2s video clips from the LRS2 dataset [2] and preprocess them following the procedure described in Sec. 5.1 in the main paper. Note that the LRS2 dataset contains only real videos. Each clip is converted into 16 cropped face frames, which are then resized to 224×224 to serve as the visual input. For the audio stream, we extract 128-dimensional log Mel filterbank features using a 25 ms Hanning window and a 10 ms hop length. The resulting spectrogram is subsequently resized to 1024×128 and used as the audio input. Both audio and visual inputs are normalized and then tokenized.

Token Masking. In the Holistic Coherence Priors Pre-training phase, we adopt the MAE framework [17] to enable efficient self-supervised pre-training, in which a large proportion of tokens are masked. For visual tokens, we deploy tube masking [43] with a ratio of 90%, where patches at the same spatial location across each frame share the same mask to reduce temporal information leakage. As for audio tokens, random masking [18] with a ratio of 81.25% is applied to achieve diverse time–frequency coverage. The masking ratios are chosen empirically based on previous research [18, 43].

Model Architecture. We adopt a symmetric architecture for both audio and visual modalities. Each encoder consists of $N_e = 12$ Transformer layers, with 12 attention heads per layer and an embedding dimension of 768. Hierarchical features are extracted from $H = 4$ layers, uniformly selected from the 3rd, 6th, 9th, and 12th layers of the encoders. The audio-visual interaction module comprises an 8-head cross-attention layer followed by an 8-head Transformer block. The cross-modal semantic decoders include a linear projection layer followed by a single Transformer block with 12 attention heads. The modality-specific decoders consist of $N_d = 4$ Transformer layers with 6 attention heads per layer and an embedding dimension of 384, four 6-head cross-attention layers corresponding to the hierarchical feature layers, and four linear heads that project each layer's output back to the original input.

Training Configuration. Following [34], we initialize the audio encoder–decoder with pretrained AudioMAE [18] weights from AudioSet-2M [14], and initialize the visual encoder–decoder with MARLIN [7] pretrained on the YouTubeFace dataset [45]. The Fine-grained Contrastive loss weight is set empirically to $\lambda_{cl} = 0.01$, and the temperature parameter τ is fixed at 0.07. The Cross-modal Semantic Reconstruction loss weight is fixed at 1. Both $\mathcal{L}_{rec}$ and $\mathcal{L}_{cross}$ are computed by averaging squared errors over the tokens. Subsequently, we pre-train the HAVIC using the AdamW optimizer [29] with a learning rate of 1.5e-4 with a cosine decay [28]. We train for 200 epochs with a linear warmup for 20 epochs using four NVIDIA L20 GPUs with a total batch size of 112. It takes about five days to complete the Holistic Coherence Priors Pre-training phase.

B.2. Holistic Adaptive Aggregation Classification

Inputs. The inputs for the Holistic Adaptive Aggregation Classification phase are drawn from the FakeAVCeleb [21] dataset, which contains deepfake videos with manipulated audio, visual, or both modalities. The processing procedure follows that of the Holistic Coherence Priors pre-training phase, with the difference that no masking is applied to the input in this phase. Following [34], we apply weighted sampling to alleviate class imbalance between real and fake samples in the FakeAVCeleb dataset.

Model Architecture. We remove the decoders of HAVIC. The audio and visual encoders, along with the audio-visual interaction module, retain the same structure as in the Holistic Coherence Priors pre-training phase. Each scoring network in the Adaptive Feature Aggregation module is a 2-layer MLP, and each classification head is a 3-layer MLP.

Training Configuration. We initialize the audio encoder, visual encoder, and audio-visual interaction module using the weights obtained from the Holistic Coherence Priors Pre-training phase. The model is then trained using the AdamW optimizer [29] with a cosine annealing scheduler with warm restarts [28]. A smaller learning rate of 1.0e-5 is applied to the pretrained components, while newly added modules are trained with a larger learning rate of 1.0e-4. Training is performed for 50 epochs with a total batch size of 32 on four NVIDIA L20 GPUs, taking about eight hours.

Figure 1. **Examples of real–fake video pairs generated by the six models in HiFi-AVDF.** For each model, we display one representative pair, where the top row shows a real video clip and the bottom row shows the corresponding forged clip produced by that model.

Inference. During inference, we follow [34] and apply a sliding-window strategy for video-level detection. Each window has a duration of 3.2s and slides with a step size of 0.4s. The output logits from the main classification head are computed for each window, and the final prediction (real or fake) is obtained by averaging the logits across all windows.

C. Dataset Details

C.1. HiFi-AVDF Dataset

Overall Dataset Creation Pipeline. After collecting the real data and extracting three core components, namely the reference frame, audio track, and video caption, as described in Sec. 4 in the main paper, forged videos are generated following three strategies to simulate diverse audio-visual manipulations:
(i) **caption-only**: Video generation is driven entirely by a text caption, creating a purely text-to-video (T2V) generation without audio-visual references.
(ii) **caption + reference frame**: A reference frame is incorporated with the text caption to ground the generated content, significantly improving visual realism.
(iii) **audio track + reference frame**: The generator synchronizes the lip movements of a subject in a reference frame with a source audio track, producing realistic, audio-visually coherent forgeries.

In practice, Sora 2 follows the caption-only strategy, as it does not support using a reference face image for video generation. Seedance 1.0 employs the audio track + reference frame strategy, while the remaining four models utilize the caption + reference frame strategy. As shown in Fig. 1, for each real video, a corresponding forged video is generated using one of the models, resulting in a dataset comprising 1,905 real videos and 1,905 corresponding fake videos.

Comparison with Existing Datasets. We compare HiFi-AVDF with representative deepfake datasets in terms of manipulated modality, dataset curation, availability of text-to-video (T2V) and image-to-video (I2V) samples, generation methods, number of persons, and counts of real and fake samples. As summarized in Tab. 1, HiFi-AVDF provides high-quality audio-visual forgeries generated using diverse state-of-the-art methods, covering a larger number of persons and offering both T2V and I2V capabilities. This comprehensive design enables more robust evaluation of audio-visual deepfake detection models and facilitates research on multi-modal forgery scenarios.

Ethical and Bias Issues. We acknowledge that the HiFi-AVDF dataset may raise ethical concerns, particularly regarding the potential misuse of facial videos and the advanced audio-visual generation tools employed in constructing the dataset. Such misuse could involve the creation of new deepfake content or other forms of malicious exploitation. To mitigate these risks, we release the dataset under a carefully designed end-user license agreement that explicitly restricts the use of the dataset and any generated audio-visual content to research purposes only. The dataset is provided solely to support scientific progress in deepfake detection, and any attempt to employ the data for harmful

Dataset	Manipulated Modality	T2V	I2V	Generation Method	Person #	Real #	Fake #	Year
FaceForensics++ [38]	V	✗	✗	FaceSwap (2017) [24], DeepFakes (2017)[1], Face2Face (2016)[41], NeuralTextures (2019)[42]	N/A	1,000	4,000	2019
WildDeepfake [50]	V	✗	✗	N/A	N/A	3,805	3,509	2020
KoDF [26]	V	✗	✗	FaceSwap [24] (2017), DeepFaceLab [36](2020), FOMM [39] (2019) FaceSwapGAN [33] (2019), ATFHP [47] (2020), Wav2Lip [37] (2020)	403	62,166	175,776	2021
DF-Platter [32]	V	✗	✗	FaceSwap (2017) [24], FaceSwapGAN (2019) [33], Faceshifter (2019) [27]	454	133,260	132,496	2023
FakeAVCeleb [21]	AV	✗	✗	Wav2Lip (2020) [37] , FaceSwap (2017) [24], FaceSwapGAN (2019) [33], SV2TTS (2018) [20]	500	500	19,500	2021
DefakeAVMiT [46]	AV	✗	✗	FaceSwap (2017) [24], Voice Replay (2017) [23], SV2TTS (2018) [20], DeepFaceLab (2020) [36], Wave2Lip (2020) [37], PC-AVS (2021) [48], EVP (2021) [19], AV exemplarAE (2020) [12]	86	540	6,480	2023
AV-Deepfake1M [8]	AV	✗	✗	VITS (2021) [22], YourTTS (2022) [9], TalkLip (2023) [44]	2068	286,721	860,039	2024
HiFi-AVDF (ours)	AV	✓	✓	Sora 2 (2025) [35], Veo 3.1 (2025) [16], Seedance 1.0 (2025) [6], Kling 2.5 (2025) [25], WAN 2.5 (2025) [4], PixVerse V5 (2025) [3]	1905	1,905	1,905	2025

Table 1. **Summary of representative deepfake datasets.** This table compares commonly used visual and audio-visual deepfake datasets in terms of manipulated modality, support for T2V/I2V generation, generation methods, number of persons, and counts of samples. HiFi-AVDF provides high-quality audio-visual forgeries generated by diverse state-of-the-art models, supporting both T2V and I2V modalities.

or non-research activities is strictly prohibited.

Furthermore, to conduct a comprehensive bias assessment of HiFi-AVDF, we perform an automated demographic analysis using EasyFace [5], categorizing individuals by binary gender, seven racial/ethnic groups, and nine age ranges spanning from infancy to older adulthood. This procedure provides a fine-grained understanding of the dataset's demographic composition. A face detection pipeline equipped with pre-trained multi-attribute recognition models was applied to all samples, and the resulting statistics are summarized in Tab. 2. While our dataset provides broad demographic coverage, it still contains certain degrees of demographic bias. Additionally, automated demographic classification may be inaccurate for edge cases and intersectional identities.

C.2. Other Datasets

LRS2 [2]. LRS2 is a large-scale, unconstrained audio-visual dataset for speech recognition. It comprises 97k real videos sourced from British television, each paired with its corresponding audio track, enabling the modeling of both real human facial movements and their corresponding audio signals, and capturing the intrinsic audio-visual coherence.

FakeAVCeleb [21]. FakeAVCeleb is a deepfake detection dataset of 20,000 videos, comprising 500 real videos from VoxCeleb2 [10] and 19,500 deepfakes created via visual (FaceSwap [24], FSGAN [33], Wav2Lip [37]) and audio (SV2TTS [20]) manipulations. Based on which modalities are manipulated, the dataset can be categorized into four types: real video with fake audio (RVFA), fake video with real audio (FVRA), fake video with fake audio (FVFA), and real video with real audio (RVRA, i.e., the unaltered samples). Furthermore, the forgeries are generated using different combinations of manipulation techniques, as summarized in Tab. 3.

KoDF [26]. KoDF is a large-scale talking-face deepfake dataset comprising 62,166 real videos and 175,776 fake videos generated using six synthesis algorithms:

Category	**Attribute**	**Percentage (%)**
Gender	Male	62.99
	Female	37.01
Race/Ethnicity	White	56.43
	Black	3.73
	Latino Hispanic	4.41
	East Asian	11.86
	Southeast Asian	1.89
	Indian	4.20
	Middle Eastern	17.48
Age	0-2	0.05
	3-9	0.79
	10-19	4.67
	20-29	37.95
	30-39	25.83
	40-49	17.90
	50-59	8.45
	60-69	3.94
	70+	0.42

Table 2. Demographic distribution analysis results on the HiFi-AVDF dataset.

Category	Generation Method
RVFA	SV2TTS
FVRA-FS	FaceSwap
FVRA-GAN	FaceSwapGAN
FVRA-WL	Wav2Lip
FVFA-FS	SV2TTS + FaceSwap
FVFA-GAN	SV2TTS + FaceSwapGAN
FVFA-WL	SV2TTS + Wav2Lip

Table 3. Overview of generation methods corresponding to different audio-visual manipulation categories in FakeAVCeleb.

Method	RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
AV-DFD [49]	74.9	73.3	97.0	97.4	99.6	99.7	58.4	55.4	**100.**	**100.**	86.0	85.2
AVAD (LRS2) [13]	62.4	71.6	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	87.8	89.9
AVAD (LRS3) [13]	70.7	80.5	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	87.2	90.3
AVFF [34]	93.3	92.4	94.8	98.2	**100.**	**100.**	99.9	**100.**	99.4	99.8	97.5	98.1
AVPrompt [31]	97.1	95.5	99.9	99.9	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	99.4	99.1
HAVIC (Ours)	**98.6**	**96.7**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**99.7**	**99.3**

Table 4. **Cross-manipulation generalization on FakeAVCeleb.** We evaluate the model on unseen manipulation types by training on four categories and testing on the held-out category. HAVIC consistently achieves superior performance across all manipulation types, demonstrating strong generalization to unseen deepfake generation methods.

FaceSwap [24], DeepFaceLab [36], FaceSwapGAN [33], FOMM [39], ATFHP [47], and Wav2Lip [37]. Following [13, 34], we use a subset of KoDF to evaluate the cross-dataset generalization (Tab. 3 in the main paper).

D. Additional Results

D.1. Cross-Manipulation Generalization.

In addition to the intra-dataset evaluation reported on the FakeAVCeleb dataset (Tab. 2 in the main paper), we further conduct cross-manipulation experiments on the FakeAVCeleb dataset, following the protocols in [13, 31, 34]. The dataset is divided into five categories according to the specific deepfake generation algorithms: RVFA, FVRA-WL, FVFA-FS, FVFA-GAN, and FVFA-WL. In each experiment, we hold out one category for testing while training the model on the remaining four categories. This leave-one-type-out evaluation setup enables us to measure the model's ability to detect unseen manipulation methods. The results are summarized in Tab. 4. HAVIC consistently outperforms all baseline methods across all manipulation types. This demonstrates that our model effectively captures generalizable audio-visual cues that transfer well to unseen deepfake generation methods.

D.2. Ablation on the absence of the audio modality.

In practical scenarios, many videos may contain no audio or severely corrupted audio tracks. Most existing audio–visual detection methods heavily rely on both modalities, making them vulnerable when audio is missing. To evaluate the robustness of HAVIC under such conditions, we perform an ablation where the audio modality is entirely removed during testing. In this setting, we only use HAVIC's visual classification head for inference. For the two compared methods [11, 34] that do not support audio-less input, we provide a silent audio track as a placeholder. As shown in Tab. 5, all compared methods experience a notable drop in performance when audio is absent. In contrast, HAVIC maintains strong performance, showing that it can still perform reliable detection using only visual information.

D.3. Comparisons with other SSL methods.

To further validate the effectiveness of our self-supervised learning (SSL) design in the Holistic Coherence Priors pre-training phase, we compare HAVIC with representative SSL methods from two perspectives.

MAE-based Pre-training. Since our method introduces hierarchical decoding and layer-wise supervision design beyond standard MAE, we compare HAVIC with several MAE variants to assess the benefit of these improvements. Specifically, we substitute our design with two representative variants: VideoMAE [43] and HiCMAE [40], and pre-train the model under the same settings. As shown in Tab. 6, VideoMAE only uses the features from the last encoder layer for reconstruction, resulting in moderate performance. HiCMAE improves upon this by incorporating skip connections between the encoder and decoder, encouraging intermediate layers to learn more meaningful representations, which leads to better performance. Building on this, our method further introduces hierarchical decoding and layer-wise supervision, yielding additional performance gains.

Audio-Visual Contrastive Learning. To evaluate the contribution of our fine-grained audio-visual contrastive learning, we conduct two ablation studies targeting its key components: temporal segmentation and the soft negative mechanism. Previous contrastive methods [15, 34] treat an entire

Method	FakeAVCeleb		KoDF		HiFi-AVDF	
	ACC	AUC	AP	AUC	AP	AUC
AVFF [34]	84.9	86.5	77.2	78.4	56.2	55.3
PIA [11]	87.2	90.3	82.1	85.9	53.7	55.8
Visual Cls. of HAVIC	96.6	98.1	91.9	93.7	59.2	61.4
HAVIC (Ours)	**99.8**	**99.9**	**99.2**	**98.9**	**75.4**	**75.7**

Table 5. Ablation on the absence of the audio modality.

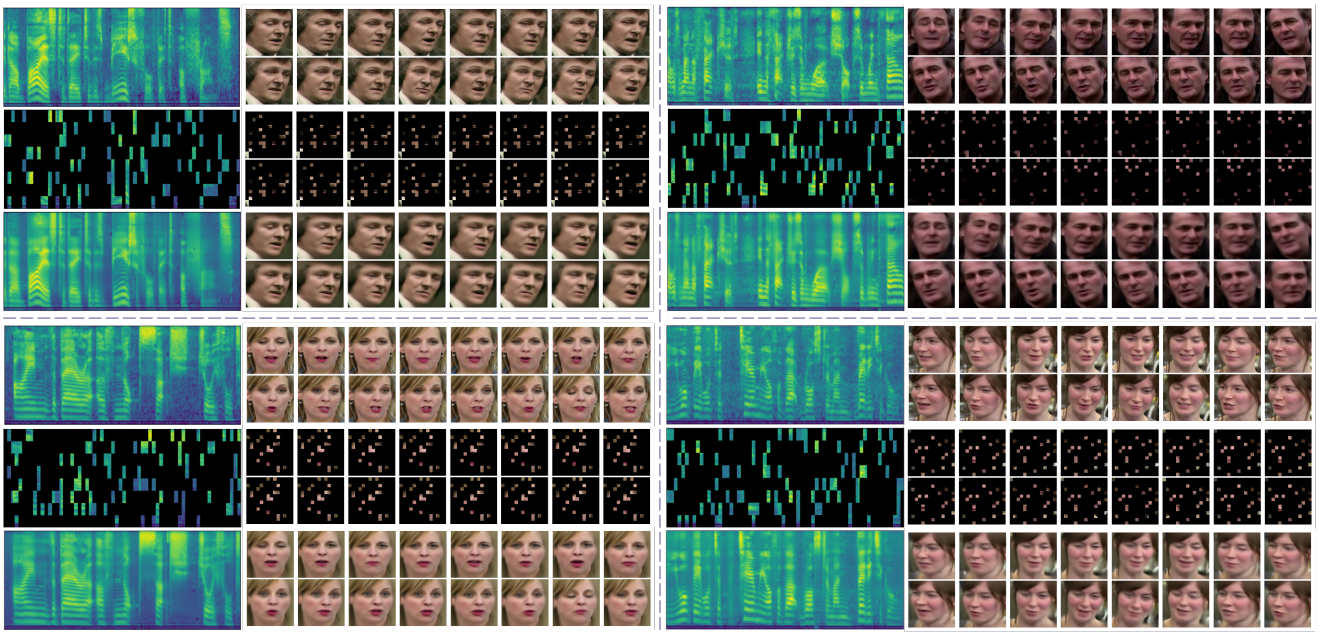

Figure 2. **Modality-Specific Hierarchical Reconstruction visualizations**. For each clip, the first row shows the original audio spectrograms and visual frames, while the second and third rows depict the masked inputs and the corresponding reconstructions from the final decoder layer, respectively. Details of the reconstructions can be seen by zooming in.

audio–video pair as a single unit, ignoring temporal structure. We first replace our segment-level formulation with a video-level contrastive loss by globally pooling both modalities to assess the impact of temporal segmentation. Second, we remove the soft negative mechanism. As shown in Tab. 7, both components contribute to the model's performance, demonstrating the effectiveness of our designs.

D.4. Hyperparameter Sensitivity of Loss Weight.

To assess the influence of the loss weight in the Fine-grained Audio-Visual Contrastive Loss, we experiment with different weighting values. As shown in Tab. 8, a weight around $\lambda_{cl} = 0.01$ yields the best performance across all datasets. A smaller weight weakens the contrastive learning signal, making the model insufficiently align audio–visual features, whereas an excessively large weight overwhelms other objectives and disrupts the overall optimization balance.

D.5. Analysis of Model Complexity.

We analyze the trade-off between model performance and computational efficiency. Tab. 9 compares the number of parameters, throughput, and performance on HiFi-AVDF dataset of representative MAE-based pre-training models. Our proposed HAVIC achieves the highest AP and AUC while maintaining competitive throughput and a moderate number of parameters, demonstrating an efficient balance between accuracy and computational cost.

Method	FakeAVCeleb		KoDF		HiFi-AVDF	
	ACC	AUC	AP	AUC	AP	AUC
VideoMAE [43]	98.4	99.1	94.5	95.2	69.1	71.3
HiCMAE [40]	99.3	99.5	96.7	96.3	71.8	73.5
HAVIC (Ours)	**99.8**	**99.9**	**99.2**	**98.9**	**75.4**	**75.7**

Table 6. Comparison with MAE-based Pre-training Variants.

Method	FakeAVCeleb		KoDF		HiFi-AVDF	
	ACC	AUC	AP	AUC	AP	AUC
w/o temporal segments	98.7	98.9	96.5	96.8	69.6	70.8
w/o soft negative mechanism	99.4	99.5	98.4	98.1	74.0	75.1
HAVIC (Ours)	**99.8**	**99.9**	**99.2**	**98.9**	**75.4**	**75.7**

Table 7. Ablation study on key components of Fine-grained Audio-Visual Contrastive Learning.

Loss weight λ_{cl}	FakeAVCeleb		KoDF		HiFi-AVDF	
	ACC	AUC	AP	AUC	AP	AUC
0.001	99.8	99.8	97.3	96.7	72.3	67.1
0.005	99.8	99.9	98.5	99.0	74.7	74.4
0.01	99.8	99.9	99.2	98.8	75.4	75.7
0.05	99.8	99.8	98.3	98.4	73.9	75.5
0.1	99.6	99.7	97.8	97.9	72.4	73.8
0.5	99.3	99.8	96.5	96.7	72.8	71.6
1	98.7	99.1	96.2	96.9	70.2	69.8

Table 8. Ablation study on loss weight for Fine-grained Audio-visual Contrastive Loss.

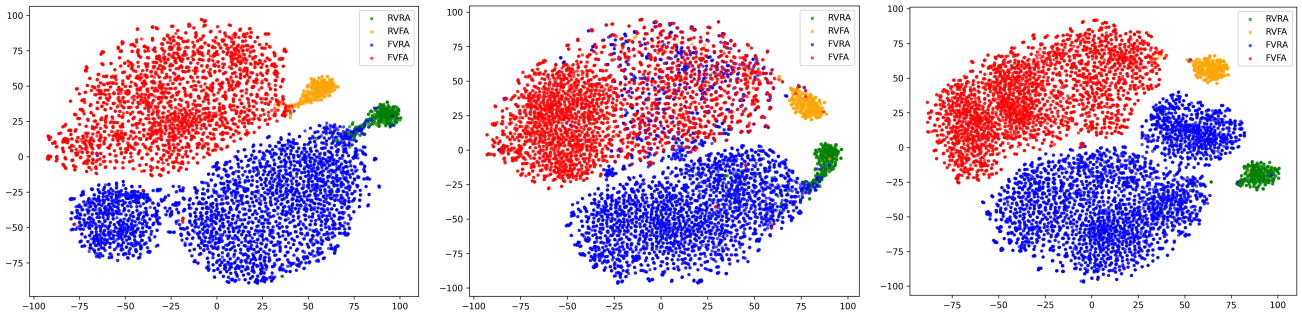

Figure 3. **t-SNE visualization of the learned audio-visual embeddings**. From left to right: (1) embeddings without Adaptive Aggregation, (2) embeddings without Auxiliary Classifiers, and (3) embeddings of the complete model.

Method	Parameters (M)	Throughput (samples/s)	AP	AUC
VideoMAE [43]	235.1	74.7	69.1	71.3
HiCMAE [40]	241.0	72.5	71.8	73.5
AVFF [34]	196.8	34.2	66.0	65.2
HAVIC (Ours)	243.1	66.8	75.4	75.7

Table 9. Comparison of model complexity and performance. The number of parameters and inference throughput are reported for the pre-training models, while AP and AUC are evaluated on the HiFi-AVDF dataset to assess detection performance.

Furthermore, we conduct an ablation study on the components of Hierarchical Adaptive Aggregation Classification phase to examine their impact on model complexity and efficiency. Tab. 10 shows that both adaptive aggregation and auxiliary classifiers contribute to improved performance, with a slight reduction in throughput as more components are added. This analysis highlights the trade-off between incorporating advanced modeling components and maintaining efficient inference speed.

Adaptive Aggregation	Auxiliary Classifiers	Parameters (M)	Throughput (samples/s)	AP	AUC
✗	✗	208.9	67.0	65.3	67.6
✓	✗	209.9	65.4	67.7	71.8
✗	✓	213.7	64.2	71.2	72.9
✓	✓	214.7	60.8	75.4	75.7

Table 10. Impact of components in the Hierarchical Adaptive Aggregation Classification phase on model size, throughput, and performance on the HiFi-AVDF dataset. ✗and ✓indicate the exclusion and inclusion of each component, respectively.

D.6. Multiple Runs

In the main paper, we report performance metrics averaged over multiple runs with different random seeds to mitigate randomness and enable reliable comparisons across three benchmark datasets. To provide a detailed view of variability, Tab. 11 presents the results of five individual runs for each dataset, along with their mean and standard deviation.

D.7. Qualitative Analysis

Modality-Specific Reconstruction Visualizations. We present visualizations of the Modality-Specific Hierarchical Reconstruction in Fig. 2. Video clips are randomly selected from the unseen test set. For each clip, the first row shows the original audio spectrogram and 16 visual frames, while the second and third rows depict the masked inputs and the corresponding reconstructions, respectively. All reconstruction results are produced by the final layer of the decoders.

The reconstructions closely resemble the original inputs, successfully recovering the overall structure of both modalities. Although the reconstructions are slightly smoother than the ground truth as a result of the high masking ratio, HAVIC still restores informative and coherent patterns. These results indicate that HAVIC can effectively infer missing content from limited visible information while capturing meaningful features, demonstrating that it has learned robust intra-modal structural coherence priors that benefit downstream audio-visual deepfake detection.

Visualization of Embedding Space with t-SNE. To further analyze the impact of the Hierarchical Adaptive Aggregation Classification phase (Tab. 6 in the main paper), we visualize the learned audio-visual embeddings using t-SNE [30]. As shown in the Fig. 3, in the leftmost plot (w/o Adaptive Aggregation), the clusters exhibit some over-

Multiple Runs	FakeAVCeleb		KoDF		HiFi-AVDF	
	ACC	AUC	AP	AUC	AP	AUC
(i)	99.81	99.95	99.32	98.86	75.93	76.04
(ii)	99.84	99.96	99.39	99.24	75.25	75.85
(iii)	99.90	99.99	98.96	98.63	75.06	74.91
(iv)	99.86	99.98	98.79	98.87	74.19	75.58
(v)	99.79	99.93	99.59	99.10	76.62	76.49
Mean	99.84	99.96	99.21	98.94	75.41	75.77
std	0.04	0.02	0.33	0.24	0.92	0.59

Table 11. Performance across 5 runs on three benchmark datasets.

lap, likely because average feature aggregation weakens the discriminative information of certain strong features. In the middle plot (w/o Auxiliary Classifiers), the overlap between forged samples is more pronounced, particularly for FVFA and FVRA, as the model is only trained to output overall real/fake predictions, limiting its ability to capture modality-specific discrepancies. In contrast, the rightmost plot shows embeddings from our full model, where the clusters are clearly separated, highlighting the effectiveness of our proposed Hierarchical Adaptive Aggregation Classification phase in producing discriminative features for deepfake detection.

References

[1] Deepfakes. https://github.com/deepfakes/faceswap. Accessed: 2025-11-03. 3

[2] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018. 1, 3

[3] AISphere. Pixverse. https://app.pixverse.ai/home. Accessed: 2025-11-03. 3

[4] Alibaba. Wan. https://tongyi.aliyun.com/wan/. Accessed: 2025-11-03. 3

[5] Sithu Aung. Easyface: Easy face analysis tool with sota models. https://github.com/sithu31296/EasyFace. Accessed: 2025-11-03. 3

[6] Bytedance. Seedance. https://seed.bytedance.com/en/seedance. Accessed: 2025-11-03. 3

[7] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1493–1504, 2023. 1

[8] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7414–7423, 2024. 3

[9] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR, 2022. 3

[10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 3

[11] Soumyya Kanti Datta, Tanvi Ranga, Chengzhe Sun, and Siwei Lyu. Pia: Deepfake detection using phoneme-temporal and identity-dynamic analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1596–1606, 2025. 4

[12] Kangle Deng, Aayush Bansal, and Deva Ramanan. Unsupervised audiovisual synthesis via exemplar autoencoders. *arXiv preprint arXiv:2001.04463*, 2020. 3

[13] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023. 4

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[15] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 4

[16] Google. Veo 3.1. https://aistudio.google.com/models/veo-3. Accessed: 2025-11-03. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[18] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35: 28708–28720, 2022. 1

[19] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 3

[20] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018. 3

[21] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 1, 3

[22] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021. 3

[23] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017. 3

[24] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 3, 4

[25] Kuaishou. Kling. https://klingai.com/. Accessed: 2025-11-03. 3

[26] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF*

international conference on computer vision, pages 10744–10753, 2021. 3

[27] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 3

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 1

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 6

[31] Hui Miao, Yuanfang Guo, Zeming Liu, and Yunhong Wang. Multi-modal deepfake detection via multi-task audio-visual prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 612–621, 2025. 4

[32] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Df-platter: Multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2023. 3

[33] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 3, 4

[34] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024. 1, 2, 4, 6

[35] OpenAI. Sora 2. https://openai.com/index/sora-2/. Accessed: 2025-11-03. 3

[36] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 3, 4

[37] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 3, 4

[38] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 3

[39] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 3, 4

[40] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382, 2024. 4, 5, 6

[41] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 3

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[43] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 1, 4, 5, 6

[44] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023. 3

[45] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011. 1

[46] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023. 3

[47] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 3, 4

[48] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 3

[49] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14800–14809, 2021. 4

[50] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2382–2390, 2020. 3