AMBIGUITY-AWARE TEMPORAL TRANSPORT FOR AUDIO-VISUAL DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning audio-visual consistency from authentic videos reduces dependence on specific manipulation patterns in deepfake detection. However, sustained or recurring visual states may produce similar matching responses across neighboring audio segments, leaving local correspondences ambiguous. Under such uncertainty, local matching strength may become less reliable as detection evidence. This motivates evaluating both correspondence informativeness and the temporal coherence of matching evidence. We propose **A**mbiguity-**A**ware **T**emporal **T**ransport (AATT), a detection framework that jointly models these two properties. AATT estimates correspondence informativeness from the normalized entropy of local matching distributions to weight evidence for temporal inference. The weighted evidence is integrated across time through marginalization over monotone paths, yielding soft correspondences between the two modalities. These correspondences guide the aggregation of audio features onto the visual timeline for an informativeness-weighted comparison with visual features. The detection score combines the resulting transport compatibility and mean correspondence informativeness with path surprisal, which measures the cost of imposing temporal constraints. The framework learns exclusively from authentic talking videos through a ranking objective that distinguishes matched audio-visual pairs from cross-video mismatches. Experiments demonstrate substantial improvements on talking benchmarks, a pronounced advantage in the more challenging singing setting, and robustness to different perturbations.

1 INTRODUCTION

Advances in generative modeling have improved the realism of audio-driven facial animation (Xie et al., 2025; Liu et al., 2025; Cui et al., 2025), making reliable audio-visual deepfake detection (Oorloff et al., 2024) more challenging. In talking and singing videos, acoustic changes and corresponding facial movements reflect a shared vocal production process (Scholes et al., 2020). The temporal relationship between these two modalities provides cross-modal evidence for assessing video authenticity. Therefore, evaluating audio-visual consistency is an important direction for detecting forged videos.

Traditional supervised audio-visual deepfake detectors (Haliassos et al., 2021; Liu et al., 2024) learn to identify cross-modal inconsistencies and manipulation artifacts from authentic and forged videos. Their reliance on generator-specific patterns may limit generalization to unseen manipulations. Recent methods instead learn audio-visual relationships from authentic videos to improve generalization without relying on forged training samples. AVAD (Feng et al., 2023) uses autoregressive modeling to identify anomalous audio-visual synchronization patterns, while AVH-Align (Smeu et al., 2025) aggregates learned frame-level alignment scores. T-AVFD (Liu et al., 2026) introduces the Singing Head DeepFake (SHDF) benchmark and reveals the challenges of generalization from talking to singing. It combines semantic representations of facial authenticity with audio-visual alignment features, adaptively weighting these complementary sources of detection evidence. However, sustained or repeated visual states in authentic audio-visual videos may produce similar matching responses across neighboring audio segments. The resulting correspondence ambiguity may reduce the reliability of local matching strength in video-level consistency assessment.

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.

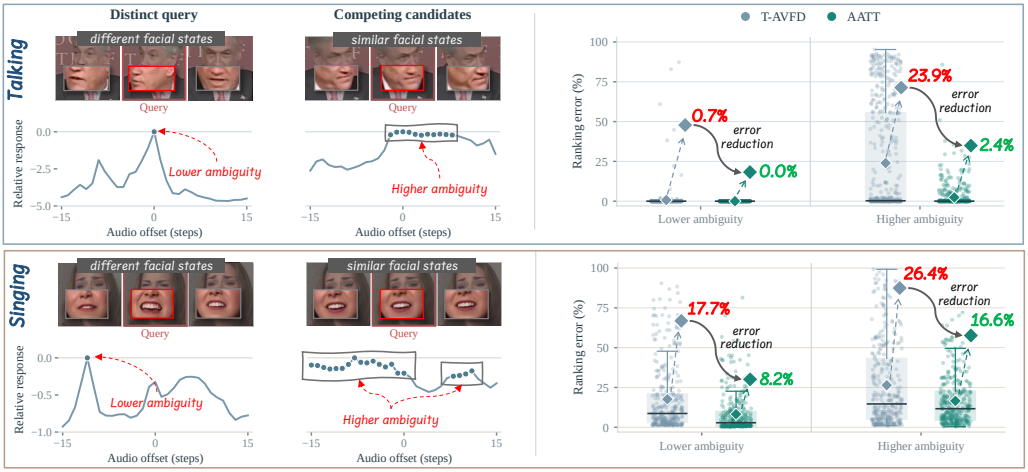

Figure 1: **Correspondence ambiguity in authentic videos and its relation to detection performance.** Red boxes mark visual queries in examples with lower or higher correspondence ambiguity. Curves show relative T-AVFD matching responses, with highlighted points indicating near-best audio candidates. The right panels compare T-AVFD and our method AATT across the two ambiguity groups in talking and singing scenarios. Only authentic videos meeting the minimum feature sequence length requirement are included in the ambiguity groups. Each dot represents the ranking error of an authentic video against manipulated videos (lower is better). Boxes span the interquartile range, internal lines mark medians, and diamonds and percentages indicate mean ranking errors.

To examine this ambiguity, we compare local matching responses for different visual queries from authentic videos in Figure 1 (left). Some queries exhibit a distinct best-matching candidate, whereas others produce similar responses across neighboring audio positions. These contrasting patterns indicate that matching strength alone does not fully capture how well local evidence distinguishes among candidate correspondences. The analysis in Figure 1 (right) relates this ambiguity to detection performance, showing higher mean ranking errors for T-AVFD in the higher ambiguity groups in both talking and singing settings. Within the same groups, our method substantially reduces these errors. These observations motivate accounting for correspondence informativeness and temporal coherence when assessing the contribution of local matching evidence to audio-visual consistency.

To this end, we propose **A**mbiguity-**A**ware **T**emporal **T**ransport (AATT), a framework that incorporates correspondence informativeness and temporal coherence into audio-visual consistency modeling. AATT assigns greater weight to local evidence that more clearly distinguishes candidate positions, then integrates this evidence across time by marginalizing over monotone paths. The resulting soft correspondences aggregate audio features across candidate positions onto the visual timeline, with correspondence informativeness further weighting the contribution of each time step to cross-modal compatibility. Beyond feature agreement, path surprisal measures the cost of imposing temporal constraints on local matching preferences, providing a complementary assessment of temporal coherence. AATT combines cross-modal transport compatibility, path surprisal, and mean correspondence informativeness to produce the final detection score. The framework learns exclusively from authentic talking videos through a ranking objective that distinguishes matched audio-visual pairs from cross-video mismatches. Experiments show that AATT outperforms the evaluated baselines on the AVLips (Liu et al., 2024), FakeAVCeleb (FKAV) (Khalid et al., 2021), and TalkingHeadBench (THB) (Xiong et al., 2026) talking datasets, achieves substantial gains on the more challenging SHDF (Liu et al., 2026) singing dataset, and remains robust under different perturbations.

Our main contributions are as follows:

- We reveal a distinction between local matching strength and correspondence informativeness in authentic audio-visual pairs. Authentic pairing does not guarantee unambiguous local correspondence, and individual matching scores do not fully capture how clearly candidate positions can be distinguished.

- We introduce AATT, a framework that jointly accounts for correspondence informativeness and temporal coherence. AATT integrates informativeness-weighted evidence over monotone temporal paths and combines transport compatibility, path surprisal, and mean correspondence informativeness for detection.
- We demonstrate substantial improvements over existing detectors across talking and singing benchmarks, with strong cross-scenario generalization, robustness to perturbations, and efficient inference. Controlled ablations further support the effectiveness of the proposed design.

2 Related Work

2.1 Audio-Visual Deepfake Detection

Audio-visual deepfake detection identifies synthesized videos by jointly analyzing audio and visual content (Ilyas et al., 2023; Miao et al., 2025; Boldisor et al., 2026; Kim et al., 2026). Advances in audio-driven facial generation (Gan et al., 2023; Xu et al., 2024b; Cheng et al., 2025; Ki et al., 2025) have made audio-visual synchronization increasingly natural, placing greater demands on detection methods. Supervised detectors learn from authentic and manipulated samples, exploiting cross-modal inconsistencies. Early methods characterize cross-modal inconsistencies through segment-level dissonance and emotional disagreement (Chugh et al., 2020; Mittal et al., 2020). Joint detection frameworks further integrate audio-visual synchronization with modality-specific detection to exploit complementary evidence (Zhou & Lim, 2021). To improve generalization, RealForensics (Haliassos et al., 2022) uses representations learned from authentic videos as auxiliary prediction targets during supervised real-fake classification. LipFD (Liu et al., 2024) focuses on inconsistencies between speech and lip movements, incorporating relationships between lip and head regions to improve lip-sync forgery detection. Despite such advances, these methods still rely on forged training samples to learn the detection boundary.

Methods trained exclusively on authentic videos instead learn natural audio-visual relationships and treat deviations from them as evidence of manipulation. Within this paradigm, AVAD (Feng et al., 2023) models synchronization features autoregressively to identify anomalous sequences, while AVH-Align (Smeu et al., 2025) aggregates learned frame-level alignment scores into video-level predictions. T-AVFD (Liu et al., 2026) combines facial authenticity semantics with alignment features through adaptive weighting and introduces the SHDF benchmark to examine generalization from talking to singing. However, authentic audio-visual pairs can still exhibit ambiguous local correspondences, complicating the use of local evidence for video-level consistency assessment. AATT accounts for this ambiguity by incorporating correspondence informativeness and temporal coherence into the evaluation of audio-visual evidence.

2.2 Audio-Visual Correspondence and Temporal Alignment

Audio-visual correspondence learning establishes relationships between acoustic signals and visual motion through representation learning and temporal matching (Gong et al., 2022). Joint embeddings enable speech and mouth movements to be compared for synchronization estimation (Chung & Zisserman, 2016), while masked multimodal prediction learns speech representations from paired audio-visual sequences (Shi et al., 2022). Beyond representation learning, explicit alignment methods determine how positions in one modality correspond to those in the other. AlignNet (Wang et al., 2020) learns dense correspondences to accommodate nonuniform temporal misalignment. Synchformer (Iashin et al., 2024) addresses synchronization with sparse informative events by separating segment-level contrastive pretraining from temporal offset estimation.

Temporal constraints further allow local matches to be evaluated jointly across a sequence (Xu et al., 2023). Soft-DTW (Cuturi & Blondel, 2017) smooths the minimum over alignment costs to obtain a differentiable sequence comparison objective, while differentiable dynamic programming incorporates structured alignment operations into neural networks (Mensch & Blondel, 2018; Dvornik et al., 2021; Zeitler & Müller, 2026). These formulations connect local matching evidence with sequence-level correspondence structure. For deepfake detection, the inferred correspondences provide a basis for assessing consistency between audio and visual sequences. AATT uses these correspondences

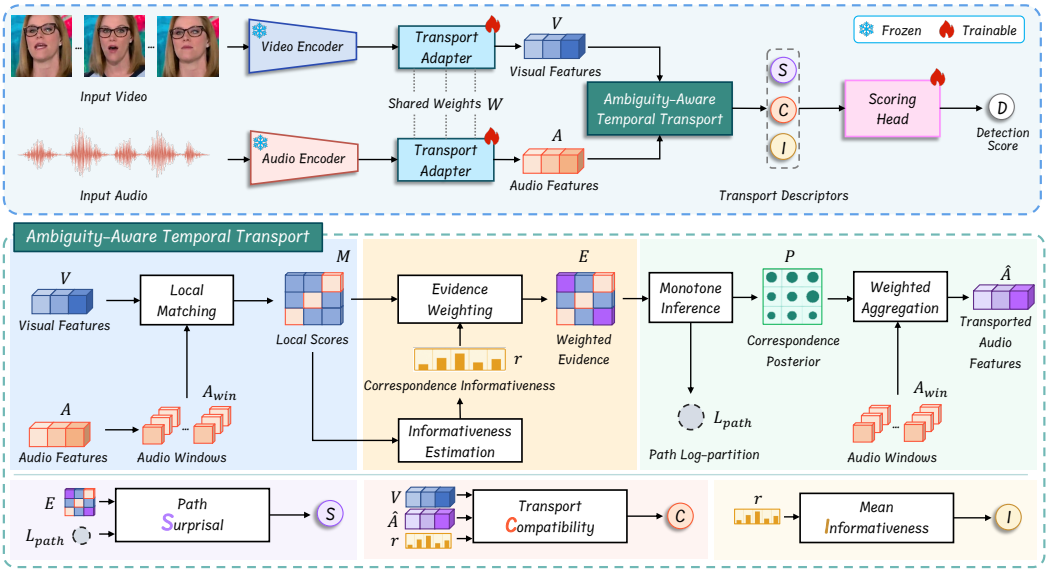

Figure 2: **Overview of AATT.** The upper panel presents the overall detection pipeline, while the lower panel details temporal transport and descriptor construction. Correspondence informativeness r weights the matching evidence for monotone path inference. The resulting soft correspondence posterior P guides the aggregation of audio features onto the visual timeline. The three colored branches produce path surprisal S, transport compatibility C, and mean correspondence informativeness I for the scoring head. The dashed connections indicate shared adapter weights.

to evaluate transport compatibility and complements this evidence with path surprisal and mean correspondence informativeness.

3 METHOD

To assess audio-visual consistency under ambiguous local correspondences, we propose Ambiguity-Aware Temporal Transport (AATT), a framework that accounts for correspondence informativeness and temporal coherence. As illustrated in Figure 2, frozen audio and visual encoders (Shi et al., 2022) extract feature sequences for the two modalities. A linear adapter with shared trainable weights W maps the L_2-normalized feature vectors to visual and audio features $V, A \in \mathbb{R}^{T\times d}$, where T denotes the common sequence length and d the feature dimension. AATT weights local matching evidence according to correspondence informativeness r and marginalizes over monotone temporal paths to obtain a soft correspondence posterior P. This posterior guides the aggregation of audio features onto the visual timeline, yielding transported features $\widehat{A} \in \mathbb{R}^{T\times d}$ for informativeness-weighted cross-modal comparison. A scoring head combines path surprisal S, transport compatibility C, and mean correspondence informativeness I into a video-level detection score D, with higher scores indicating stronger evidence of manipulation. The adapter and scoring head are trained exclusively on authentic talking videos using a ranking loss $\mathcal{L}_{\text{rank}}$ that distinguishes matched audio-visual pairs from cross-video mismatches.

3.1 CORRESPONDENCE INFORMATIVENESS

To characterize local correspondence ambiguity, we examine how matching evidence is distributed across neighboring audio positions. Let V_t and A_t denote the feature vectors at time step t in V and A, respectively, for $t = 1, \ldots, T$. For each visual query V_t, we consider audio candidates $A_{t+\delta}$ within a temporal window of radius τ, where δ is an integer offset in feature time steps. The valid offsets form the set $\mathcal{D}_t = \{\delta \in \mathbb{Z} \mid |\delta| \le \tau,\ 1 \le t+\delta \le T\}$, and $K_t = |\mathcal{D}_t|$ denotes the number of valid candidates. We consider $T \ge 2$ and an integer window radius $\tau \ge 1$, so that $K_t > 1$. These candidates form the local audio windows A_{win}. Each valid entry $M_{t\delta}$ of the local matching score

matrix M measures the cosine similarity between V_t and $A_{t+\delta}$:

$$M_{t\delta} = \cos(V_t, A_{t+\delta}), \qquad \delta \in \mathcal{D}_t. \tag{1}$$

We estimate correspondence informativeness r from the relative matching preferences in M. To control differences in score scale across queries, we standardize the valid responses using their mean μ_t and population standard deviation σ_t. For $\delta \in \mathcal{D}_t$, the standardized scores and local matching probabilities are:

$$\widetilde{M}_{t\delta} = \mathrm{clip}\left(\frac{M_{t\delta} - \mu_t}{\sigma_t + \epsilon}, -\kappa, \kappa\right), \qquad q_{t\delta} = \frac{\exp(\widetilde{M}_{t\delta})}{\sum_{\delta' \in \mathcal{D}_t} \exp(\widetilde{M}_{t\delta'})}, \tag{2}$$

where $\epsilon > 0$ ensures numerical stability and clip restricts its input to $[-\kappa, \kappa]$ for a fixed threshold $\kappa > 0$. The resulting distribution $q_t = (q_{t\delta})_{\delta \in \mathcal{D}_t}$ captures local matching preferences before temporal constraints are imposed. We quantify correspondence informativeness using the complement of the normalized entropy of q_t. With $H(q_t) = -\sum_{\delta \in \mathcal{D}_t} q_{t\delta} \log q_{t\delta}$ denoting its entropy, the correspondence informativeness is:

$$r_t = 1 - \frac{H(q_t)}{\log K_t}, \qquad r = (r_1, \ldots, r_T)^\top. \tag{3}$$

Normalization by $\log K_t$ accounts for the number of valid candidates and gives $r_t \in [0, 1]$. Larger values indicate more concentrated local matching preferences. r regulates the contribution of local evidence to temporal inference and cross-modal comparison after transport.

3.2 Monotone Temporal Transport

Using correspondence informativeness r, we modulate the standardized scores $\widetilde{M}$ to form the weighted evidence matrix E for joint temporal inference:

$$E_{t\delta} = r_t \widetilde{M}_{t\delta}, \qquad \delta \in \mathcal{D}_t. \tag{4}$$

Each entry $E_{t\delta}$ of E represents the weighted evidence for matching V_t with $A_{t+\delta}$. For less informative queries, this modulation attenuates local score differences before the evidence is integrated across time. A correspondence path $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_T)$ selects audio position $t + \delta_t$ for each visual time step t. To preserve temporal order while allowing local variations in correspondence, we restrict consecutive offsets to differ by at most one. The admissible path set is:

$$\Omega = \{\boldsymbol{\delta} \in \mathcal{D}_1 \times \cdots \times \mathcal{D}_T \mid |\delta_t - \delta_{t-1}| \le 1,\ t = 2, \ldots, T\}. \tag{5}$$

The associated audio index $t + \delta_t$ is nondecreasing, with permitted increments of zero, one, or two steps as visual time advances. Monotone inference integrates the weighted evidence E over the admissible path set Ω, yielding the correspondence posterior P and the path log-partition L_{path}. The path log-partition and the associated path distribution π are defined as:

$$L_{\mathrm{path}} = \log \sum_{\boldsymbol{\delta} \in \Omega} \exp\left(\sum_{t=1}^{T} E_{t\delta_t}\right), \qquad \pi(\boldsymbol{\delta} \mid E) = \exp\left(\sum_{t=1}^{T} E_{t\delta_t} - L_{\mathrm{path}}\right). \tag{6}$$

Marginalizing this distribution yields the correspondence posterior P. Each entry $P_{t\delta}$ gives the probability of selecting offset δ at time step t under the path distribution:

$$P_{t\delta} = \sum_{\substack{\boldsymbol{\delta}' \in \Omega \\ \delta'_t = \delta}} \pi(\boldsymbol{\delta}' \mid E), \qquad \delta \in \mathcal{D}_t. \tag{7}$$

The posterior satisfies $\sum_{\delta \in \mathcal{D}_t} P_{t\delta} = 1$, and each marginal incorporates evidence from paths spanning the sequence. Forward-backward inference computes P, with L_{path} obtained from the forward recursion. The correspondence posterior P guides the aggregation of candidates in the local audio windows A_{win}. The resulting transported audio sequence $\widehat{A}$ contains one feature for each visual time step:

$$\widehat{A}_t = \sum_{\delta \in \mathcal{D}_t} P_{t\delta} A_{t+\delta}. \tag{8}$$

This soft aggregation allows multiple audio candidates to contribute, with their relative contributions determined jointly by local evidence and temporal context. The transported audio features $\widehat{A}$ collect $\widehat{A}_t$ across visual time steps for cross-modal comparison, while E and L_{path} are used to compute path surprisal.

3.3 Consistency Assessment and Learning

Monotone inference integrates local evidence under temporal constraints. We assess how strongly the evidence supports these constraints and how well the transported audio agrees with the visual sequence. We quantify temporal support through path surprisal S, comparing the path log-partition L_{path} with an independent log-partition L_{ind} that removes constraints between time steps:

$$L_{\text{ind}} = \sum_{t=1}^{T} \log \sum_{\delta \in \mathcal{D}_t} \exp(E_{t\delta}), \qquad S = \max\left(0, \frac{L_{\text{ind}} - L_{\text{path}}}{T}\right). \tag{9}$$

Under independent candidate selection with probabilities proportional to $\exp(E_{t\delta})$, the probability of forming an admissible path is $\exp(L_{\text{path}} - L_{\text{ind}})$. Path surprisal S measures the corresponding negative log-probability per time step, with larger values indicating weaker support for temporally admissible correspondences.

Transport compatibility C complements this structural assessment by comparing the visual sequence V with the transported audio features $\widehat{A}$. Correspondence informativeness r_t weights the contribution of each time step to the pooled cosine similarity:

$$C = \frac{\sum_{t=1}^{T} r_t \cos(V_t, \widehat{A}_t)}{\max\left(\sum_{t=1}^{T} r_t, \epsilon_C\right)}, \tag{10}$$

where $\epsilon_C > 0$ ensures numerical stability. This pooling reduces the contribution of queries with diffuse local matching preferences when evaluating agreement after transport. Alongside temporal support and feature agreement, we retain mean correspondence informativeness I to summarize the concentration of local matching preferences across the sequence:

$$I = \frac{1}{T} \sum_{t=1}^{T} r_t. \tag{11}$$

A linear scoring head maps the three descriptors to a normality logit ℓ, followed by a decreasing transformation to obtain the detection score D:

$$\ell = \mathbf{w}^\top [C, -S, I]^\top + b, \qquad D = \log\left(1 + \exp(-\ell)\right), \tag{12}$$

where $\mathbf{w}$ and b are the scoring head parameters. Higher normality logits yield lower detection scores, with larger D indicating stronger evidence of manipulation.

Training uses exclusively authentic talking videos. Each visual sequence is paired with its original audio to form a matched pair and with audio from a different video in the batch to form a mismatched pair. Let ℓ_i^+ and ℓ_i^- denote the normality logits of these matched and mismatched pairs. For a batch of size B, we optimize the ranking objective:

$$\mathcal{L}_{\text{rank}} = \frac{1}{B} \sum_{i=1}^{B} \log\left(1 + \exp\left(\ell_i^- - \ell_i^+\right)\right). \tag{13}$$

This objective encourages matched pairs to receive higher normality logits than cross-video mismatches. The shared adapter and the scoring head are optimized through this objective.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate AATT on four benchmarks covering talking and singing scenarios. The talking benchmarks include AVLips (Liu et al., 2024), FakeAVCeleb (FKAV) (Khalid et al., 2021), and TalkingHeadBench (THB) (Xiong et al., 2026). For singing, we use Singing Head DeepFake (SHDF) (Liu et al., 2026), with approximately 800 authentic and 1,500 synthesized videos in the test set. AVLips is divided into training, validation, and test sets in a $6:1:3$ ratio. The FKAV test set comprises 500 authentic and 1,000 manipulated videos, while the THB test set combines all official test videos. Unless otherwise specified, AATT is trained exclusively on authentic videos from the AVLips training split and evaluated on all four benchmarks. To assess the effect of training distribution, we retrain AATT, AVH-Align and T-AVFD on authentic SHDF training videos and evaluate the resulting models on both talking and singing benchmarks.

Table 1: **Comparison with existing deepfake detectors on talking and singing benchmarks.** We report AP (%) and AUC (%), with the best results in bold and the second-best results underlined. The training data column specifies whether each detector is trained on authentic videos only or on both authentic and manipulated videos. We use released checkpoints for supervised baselines and the provided weights for AVAD, for which training code is unavailable. AVH-Align and T-AVFD are retrained on the same authentic talking data for fair comparison.

Method	Training data	**AVLips** *(talking)*		**FKAV** *(talking)*		**THB** *(talking)*		**SHDF** *(singing)*	
		AP	AUC	AP	AUC	AP	AUC	AP	AUC
CViT	Real & Fake	63.5	63.1	91.1	88.5	44.5	42.1	62.7	49.5
EfficientViT	Real & Fake	63.3	64.8	95.1	90.9	31.6	21.7	66.6	46.5
RealForensics	Real & Fake	69.9	71.9	94.2	88.2	68.7	74.3	67.7	50.9
LipFD	Real & Fake	<u>85.3</u>	84.7	83.4	77.0	45.0	49.2	38.1	50.5
AVAD	Real only	76.5	73.2	92.1	84.8	43.8	48.1	62.4	48.3
AVH-Align	Real only	74.3	84.5	93.5	93.0	64.8	82.3	55.2	37.4
T-AVFD	Real only	83.6	<u>87.7</u>	<u>95.6</u>	<u>95.6</u>	<u>87.6</u>	<u>93.0</u>	<u>85.7</u>	<u>80.2</u>
AATT	Real only	**98.9**	**98.8**	**98.8**	**98.2**	**98.9**	**99.3**	**93.2**	**88.1**

Implementation Details. We compare AATT with supervised and unsupervised deepfake detectors. The supervised baselines include CViT (Wodajo & Atnafu, 2021), EfficientViT (Coccomini et al., 2022), RealForensics (Haliassos et al., 2022), and LipFD (Liu et al., 2024). The unsupervised baselines comprise AVAD (Feng et al., 2023), AVH-Align (Smeu et al., 2025), and T-AVFD (Liu et al., 2026). We train AATT on an NVIDIA A100 GPU using AdamW (Zhou et al., 2024), with both the learning rate and weight decay set to 10^{-4}. Training uses a batch size of 8 and two random temporal crops of 96 feature time steps per video. The shared adapter weights W are initialized to the identity matrix. Performance is measured using Average Precision (AP) and Area Under the Curve (AUC). *Additional details and analyses are provided in Appendix A.*

4.2 Experimental Results

Generalization across Talking and Singing Scenarios. As shown in Table 1, AATT achieves the best AP and AUC across all four benchmarks. On the talking datasets, the gains are particularly substantial on AVLips and THB, while AATT also leads on FKAV, where several baselines already perform well. These improvements extend from the AVLips test split to external talking datasets, demonstrating the transferability of the learned consistency measure. Among the compared methods, AVAD, AVH-Align, and T-AVFD also learn exclusively from authentic talking videos. The gains over these methods support the effectiveness of jointly incorporating correspondence informativeness and temporal coherence into audio-visual consistency modeling. The advantage persists on the more challenging singing dataset SHDF, where AATT substantially outperforms existing detectors. Trained exclusively on authentic talking videos, AATT remains effective when applied to singing without further adaptation. The gains across both scenarios support our motivation to account for local correspondence ambiguity in audio-visual consistency modeling.

Table 2: **Performance after training on singing data.** All methods are trained exclusively on authentic SHDF videos. We report AP (%) and AUC (%), with the best results in bold and the second-best results underlined.

Method	**AVLips**		**SHDF**	
	AP	AUC	AP	AUC
AVH-Align	57.7	52.6	72.6	63.5
T-AVFD	<u>80.3</u>	<u>77.3</u>	<u>80.7</u>	<u>73.5</u>
AATT	**87.7**	**93.9**	**85.2**	**80.1**

Effect of the Training Distribution. To examine the influence of the training distribution, we retrain AATT, AVH-Align, and T-AVFD on authentic singing videos and evaluate them on AVLips and SHDF. As shown in Table 2, AATT achieves the best AP and AUC on both benchmarks. Compared with training on talking videos (Table 1), AVH-Align improves on SHDF but drops to a near-chance AUC on AVLips, indicating sensitivity to the training scenario. T-AVFD transfers more effectively to talking videos, while AATT achieves a further AUC gain of 16.6, indicating that its consistency measure remains highly discriminative in a vocal setting absent from training. The gains over de-

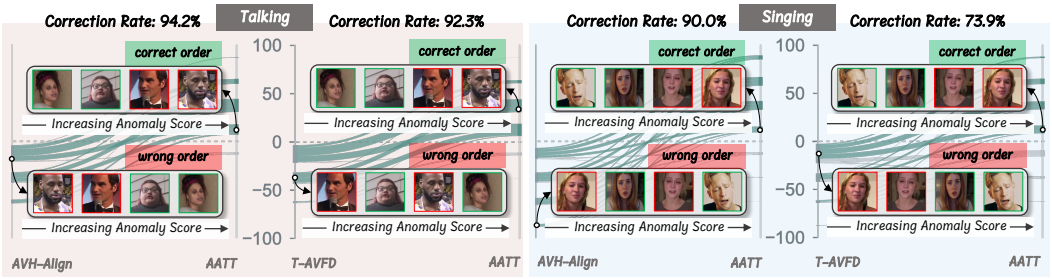

Figure 3: **Recovery of ranking errors.** Margins are fake-minus-real score percentile differences, computed separately for each model on the common test set. Positive values indicate correct ordering. Teal and gray denote recovered order and persistent reversals under AATT. Insets show baseline (bottom) and AATT (top) orderings, with green/red borders denoting real/fake videos. Anomaly scores increase from left to right, with higher scores indicating stronger evidence of manipulation. Correction rates report the fraction of baseline-misranked pairs corrected by AATT.

tectors trained on the same authentic singing data support evaluating local evidence through both correspondence informativeness and temporal coherence.

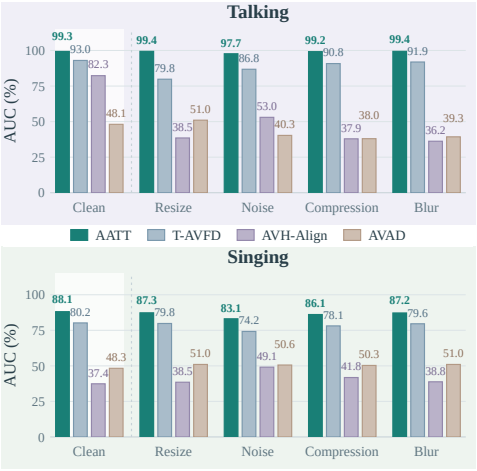

Figure 4: **Robustness to different perturbations.** AUC (%) is reported on talking (THB) and singing (SHDF) benchmarks. Colors distinguish detectors, and Clean denotes unperturbed inputs.

Recovery of Ranking Errors. To quantify the extent to which AATT corrects ranking errors, we examine its ordering of real-fake video pairs misranked by AVH-Align and T-AVFD on the talking (AVLips) and singing (SHDF) benchmarks. As shown in Figure 3, the dominant ribbons move from negative baseline margins to positive margins under AATT, indicating recovery of the correct relative ordering. The accompanying examples illustrate how reversed or interleaved real-fake orderings under the baselines become correctly ordered under AATT. Recovery exceeds 92% for both baselines on talking scenario, with substantial recovery also observed on singing scenario. These corrections show that AATT provides discriminative evidence for many video pairs that existing detectors fail to order correctly. The consistent recovery across baselines and scenarios supports the value of jointly incorporating correspondence informativeness and temporal coherence into audio-visual consistency modeling.

Robustness to Perturbations. We evaluate robustness to resize, noise, compression, and blur on talking (THB) and singing (SHDF) benchmarks. As shown in Figure 4, AATT achieves the highest AUC under all four perturbations on both benchmarks. On THB, resize, compression, and blur leave its performance close to the clean result. Noise causes the largest decrease for AATT, with AUCs of 97.7% on THB and 83.1% on SHDF. Both scores remain above those of T-AVFD on clean inputs. Perturbations may alter matching responses and make neighboring candidates harder to distinguish. Frame-level score pooling in AVH-Align and adaptive feature fusion in T-AVFD do not explicitly use correspondence informativeness to regulate the contribution of local matching evidence. The sustained advantage of AATT supports weighting local evidence by its correspondence informativeness and evaluating matching preferences under temporal constraints.

Efficiency and Performance. As shown in Table 3, AATT achieves the highest average AP and AUC with the fewest total parameters among the compared methods. It outperforms LipFD and T-AVFD while requiring substantially less computation and lower per-clip latency. Relative to T-AVFD, AATT requires 98.9% fewer GFLOPs and achieves an $11.8\times$ inference speedup. The shared linear adapter and compact scoring head contain only 1.05 M trainable parameters, while correspon-

Table 3: **Efficiency and detection performance.** Params denotes total model parameters, including frozen encoders. Average AP (%) and AUC (%) are unweighted means across AVLips, FKAV, THB, and SHDF. The best values are in bold and the second-best values are underlined.

Method	Params (M)	GFLOPs	Latency (ms/clip)	Peak mem. (GiB)	Avg. AP	Avg. AUC
LipFD	327.48	2561.96	1033.30	**1.29**	63.0	65.4
AVH-Align	<u>326.36</u>	**94.87**	**33.32**	<u>1.32</u>	72.0	74.3
T-AVFD	756.12	8281.25	724.81	1.79	<u>88.1</u>	<u>89.1</u>
AATT	**325.67**	<u>94.91</u>	<u>61.35</u>	<u>1.32</u>	**97.5**	**96.1**

dence estimation and temporal transport introduce no additional trainable parameters. Compared with AVH-Align, AATT delivers substantially stronger detection with nearly identical GFLOPs and peak memory, highlighting the effectiveness of the proposed consistency modeling while maintaining efficient inference.

4.3 Ablation Study

Table 4: **Ablation study.** Each variant is trained separately on the same authentic videos. We report AP (%) and AUC (%). The best results are bold and the second-best results are underlined.

Variant	SHDF		THB	
	AP	AUC	AP	AUC
w/o r and I	86.5	79.4	94.3	96.5
w/o r weighting	85.3	78.4	95.2	97.0
Indep. matching	<u>89.5</u>	<u>82.6</u>	<u>97.7</u>	98.3
S only	78.5	70.8	91.3	95.3
C only	84.6	75.8	92.2	93.0
I only	73.0	66.0	97.4	<u>98.5</u>
Fixed $W = \mathbf{I}_d$	82.7	75.5	85.9	94.4
AATT	**93.2**	**88.1**	**98.9**	**99.3**

Informativeness and Temporal Modeling. As shown in Table 4, removing correspondence informativeness from both evidence weighting and the detection head (w/o r and I) reduces AUC by 8.7 and 2.8 on SHDF and THB. The variant without informativeness weighting (w/o r weighting) retains I but performs substantially worse than the full model. This result highlights the value of using correspondence informativeness to regulate local evidence during temporal inference and cross-modal comparison, beyond its role as a video-level descriptor. The independent local matching variant (Indep. matching) also reduces performance, with an AUC drop of 5.5 on SHDF. This variant replaces monotone inference with independent local correspondences, for which path surprisal S becomes zero in the absence of path constraints. The comparison supports the combined value of joint temporal inference and the associated structural evidence.

Detection Descriptors. Using S, C, or I as the sole input to the scoring head reduces performance on both benchmarks. In particular, I alone achieves an AUC of 98.5% on THB but only 66.0% on SHDF, showing that the discriminative value of mean correspondence informativeness varies substantially between the two scenarios. These results support combining mean correspondence informativeness with transport compatibility and path surprisal for detection.

Learnable Feature Adaptation. Fixing the adapter weights W to the identity matrix $\mathbf{I}_d$ reduces AP by 10.5 and 13.0 on SHDF and THB, respectively. Since the scoring head remains trainable, this comparison isolates the contribution of the learned feature mapping. The reductions in both AP and AUC demonstrate the value of learning a feature space that supports correspondence modeling and discriminative audio-visual consistency assessment.

5 Conclusion

In this paper, we have examined local correspondence ambiguity in authentic audio-visual pairs, highlighting the limitations of matching strength alone for consistency assessment. We introduce AATT to jointly model correspondence informativeness and temporal coherence through evidence weighting and monotone path marginalization. Trained exclusively on authentic videos, AATT achieves substantial gains across talking and singing benchmarks, with effective cross-scenario transfer and robustness to the evaluated perturbations.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and grammar checking to improve the clarity and readability of the manuscript. We reviewed all AI-assisted text and revised it where necessary to preserve the intended technical meaning and ensure consistency with our methods and experimental results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 3 describes the proposed method, including its overall architecture and main workflow. Section 4.1 summarizes the experimental setup, with detailed settings provided in Appendix A.2.1. Appendix A.2.2 explains the principles underlying the visualizations used to motivate our method. Appendices A.2.3 and A.2.4 provide detailed information on the compared methods and the datasets used. Inference code is provided in the supplementary material to facilitate reproducibility. The complete code and model weights will be made publicly available after the review process concludes.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

Dragos-Alexandru Boldisor, Stefan Smeu, Dan Oneata, and Elisabeta Oneata. Investigating self-supervised representations for audio-visual deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43018–43029, 2026.

Hanbo Cheng, Limin Lin, Chenyu Liu, Pengcheng Xia, Pengfei Hu, Jiefeng Ma, Jun Du, and Jia Pan. Dawn: Dynamic frame avatar with non-autoregressive diffusion framework for talking head video generation. In *International Conference on Learning Representations*, volume 2025, pp. 99250–99268, 2025.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 439–447, 2020.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

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

Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pp. 219–229. Springer, 2022.

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In *International Conference on Learning Representations*, volume 2025, pp. 91659–91671, 2025.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR, 2017.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8498–8507, 2024.

Mikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan Jepson. Drop-dtw: Aligning common signal between sequences while dropping outliers. *Advances in Neural Information Processing Systems*, 34:13782–13793, 2021.

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*, pp. 10491–10503, 2023.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22577–22588. IEEE, 2023.

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.

Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5037–5047. IEEE, 2021.

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 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.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329. IEEE, 2024.

Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136:110124, 2023.

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.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

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

Taekyung Ki, Dongchan Min, and Gyeongsu Chae. Float: Generative motion latent flow matching for audio-driven talking portrait. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14699–14710. IEEE, 2025.

Youngseo Kim, Kwan Yun, Seokhyeon Hong, Sihun Cha, Colette Suhjung Koo, and Junyong Noh. X-avdt: Audio-visual cross-attention for robust deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4403–4414, 2026.

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*, pp. 3677–3685, 2017.

Ke Liu, Jiwei Wei, Shiyuan He, Zeyu Ma, Chaoning Zhang, Ning Xie, and Yang Yang. Syncgaussian: Stable 3d gaussian-based talking head generation with enhanced lip sync via discriminative speech features. In *IJCAI*, pp. 1576–1584, 2025.

Ke Liu, Jiwei Wei, Wenyu Zhang, Shuchang Zhou, Ruikun Chai, Yutao Dai, Chaoning Zhang, and Yang Yang. From talking to singing: A new challenge for audio-visual deepfake detection. In *Forty-third International Conference on Machine Learning*, 2026.

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *Advances in Neural Information Processing Systems*, 37:91131–91155, 2024.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pp. 23803–23828. pmlr, 2023.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *International Conference on Machine Learning*, pp. 3462–3471. PMLR, 2018.

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*, volume 39, pp. 612–621, 2025.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2823–2832, 2020.

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*, pp. 7184–7193, 2019.

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*, pp. 27102–27112, 2024.

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*, pp. 484–492, 2020.

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*, pp. 1–11, 2019.

Chris Scholes, Jeremy I Skipper, and Alan Johnston. The interrelationship between the face and vocal tract configuration during audiovisual speech. *Proceedings of the National Academy of Sciences*, 117(51):32791–32798, 2020.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18815–18825, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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*, pp. 14653–14662, 2023.

Jianren Wang, Zhaoyuan Fang, and Hang Zhao. Alignnet: A unifying approach to audio-visual alignment. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3298–3306. IEEE, 2020.

Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*, 2021.

Jiajian Xie, Shengyu Zhang, Mengze Li, Zhou Zhao, Fei Wu, et al. Ecoface: Audio-visual emotional co-disentanglement speech-driven 3d talking face generation. In *International Conference on Learning Representations*, volume 2025, pp. 101625–101643, 2025.

Xinqi Xiong, Prakrut Patel, Qingyuan Fan, Amisha Wadhwa, Sarathy Selvam, Xiao Guo, Luchao Qi, Xiaoming Liu, and Roni Sengupta. Talkingheadbench: A multi-modal benchmark & analysis of talking-head deepfake detection. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4139–4149. IEEE, 2026.

Ming Xu, Sourav Garg, Michael Milford, and Stephen Gould. Deep declarative dynamic time warping for end-to-end learning of alignment paths. *arXiv preprint arXiv:2303.10778*, 2023.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024a.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024b.

Johannes Zeitler and Meinard Müller. A unified perspective on ctc and soft-dtw using differentiable dtw. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8652–8661, 2023.

Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6486–6493, 2024.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.

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

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

A Appendix

A.1 Overview

This appendix presents the experimental settings, the protocol for the motivation analysis, and descriptions of the datasets and baselines. We further examine the training objective, transport direction, temporal window radius, and score distributions, and discuss directions for future research. *We provide inference code in the supplementary material and will publicly release the full code and model weights after the review process.*

A.2 More Experimental Details

A.2.1 Experimental Settings

Implementation and Training. Table 5 summarizes the configuration of AATT. The frozen audio and visual encoders produce 1,024-dimensional features. The shared adapter is initialized to the identity mapping and optimized jointly with the scoring head using the ranking objective. Training uses temporally corresponding audio and visual crops, with negative pairs constructed using audio from different videos in the batch. The window radius τ determines the local candidate range, while κ bounds the standardized matching scores. The constants ϵ and ϵ_C provide numerical stability during score standardization and compatibility aggregation.

Table 5: **Implementation and training configuration of AATT.** Temporal lengths are measured in feature time steps. The symbols correspond to those defined in the method section.

Setting	Value
GPU	NVIDIA A100 80 GB
Optimizer	AdamW
Learning rate	10^{-4}
Weight decay	10^{-4}
Batch size	8
Temporal crop length	96
Random crops per video	2
Feature dimension d	1,024
Adapter initialization	$W = \mathbf{I}_d$
Temporal window radius τ	15
Score-clipping threshold κ	4
Standardization constant ϵ	10^{-4}
Compatibility pooling constant ϵ_C	10^{-8}

Perturbation Settings. We evaluate robustness on THB and SHDF using four perturbations: resize, Gaussian noise, JPEG compression, and Gaussian blur. Each perturbation is applied separately to authentic and manipulated test videos. As summarized in Table 6, these transformations reduce spatial resolution, introduce random noise, or degrade image detail.

A.2.2 Protocol for the Motivation Analysis

The analysis in Figure 1 examines correspondence ambiguity on AVLips (talking) and SHDF (singing). We include only authentic videos with at least 80 feature time steps. We use T-AVFD to quantify ambiguity and compare both methods within the same video groups.

Local Response Analysis. We examine T-AVFD's matching responses across neighboring audio candidates. As in Section 3.1, t denotes the position of a visual query, and δ denotes the offset of an audio candidate relative to that query, measured in feature time steps. We use a window radius of $\tau = 15$ and sample 32 approximately uniformly spaced queries per authentic video. Only positions with all $K_t = |\mathcal{D}_t| = 31$ candidates available are considered, where $\mathcal{D}_t$ is the valid offset set defined in the main text. For each query, we vary the audio candidate while keeping the visual query and global video feature fixed. Let $\ell^{\mathrm{base}}_{t\delta}(v)$ denote T-AVFD's normality logit for query t and audio offset δ in authentic video v, with larger values indicating stronger matching responses. We construct the diagnostic distribution:

$$q^{\mathrm{base}}_{t\delta}(v) = \frac{\exp\big(\ell^{\mathrm{base}}_{t\delta}(v)/\sigma_{\mathrm{base}}\big)}{\sum_{\delta' \in \mathcal{D}_t} \exp\big(\ell^{\mathrm{base}}_{t\delta'}(v)/\sigma_{\mathrm{base}}\big)}. \tag{14}$$

Here, the superscript base identifies T-AVFD quantities, and σ_{base} is a fixed scale estimated from authentic validation videos. Specifically, we average candidate-logit standard deviations over queries within each validation video and take the median across videos. This fixed scaling differs from AATT's query-wise standardization, so the diagnostic distribution is distinguished from the method's local distribution q_t.

Video-level Ambiguity and Grouping. We summarize correspondence ambiguity by averaging normalized entropy over the sampled queries. Let $\mathcal{T}_v$ denote the set of 32 sampled query positions in

Table 6: **Perturbation settings for robustness evaluation.** The four perturbations are applied separately to video frames on THB and SHDF using the same parameter values. The noise standard deviation is measured on the $[0, 255]$ pixel scale.

Perturbation	Parameter	Value
Resize	Spatial scaling factor	0.5
Gaussian noise	Noise standard deviation	25
JPEG compression	Quality	20
Gaussian blur	Kernel size	5

video v, and let $q_t^{\mathrm{base}}(v) = (q_{t\delta}^{\mathrm{base}}(v))_{\delta \in \mathcal{D}_t}$ denote the corresponding candidate distribution. Using the entropy function $H(\cdot)$ defined in the main text, we define video-level ambiguity as:

$$h(v) = \frac{1}{|\mathcal{T}_v|} \sum_{t \in \mathcal{T}_v} \frac{H\big(q_t^{\mathrm{base}}(v)\big)}{\log K_t}. \tag{15}$$

Larger values of $h(v)$ indicate more diffuse local matching preferences. Within each dataset, authentic videos at or below the median ambiguity form the lower-ambiguity group, and those above it form the higher-ambiguity group. These T-AVFD-derived groups remain identical for both methods and do not depend on their detection errors.

Video-level Ranking Error. We evaluate each authentic video against the complete set of manipulated test videos from the same dataset. Extending the detection-score notation in the main text, let $D_m(v)$ denote the native video-level anomaly score assigned by method m, where m is T-AVFD or AATT and larger scores indicate stronger evidence of manipulation. For AATT, this is the score D defined in Equation 12. The ranking error of authentic video v is:

$$e_m(v) = \frac{100}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \left[\mathbf{1}\{D_m(f) < D_m(v)\} + \frac{1}{2}\mathbf{1}\{D_m(f) = D_m(v)\} \right], \tag{16}$$

where $\mathcal{F}$ is the manipulated test-video set, f denotes one of its videos, and $\mathbf{1}\{\cdot\}$ is the indicator function. This percentage counts manipulated videos incorrectly ranked as more authentic than the reference real video, with ties contributing half an error. Lower values indicate better separation. Each plotted point represents one authentic video, and group means are arithmetic averages of these errors. Both methods retain their native video-level scoring procedures, independently of the diagnostic ambiguity calculation.

A.2.3 Dataset Details

Talking Head. We evaluate AATT on three talking head benchmarks: AVLips (Liu et al., 2024), FakeAVCeleb (FKAV) (Khalid et al., 2021), and TalkingHeadBench (THB) (Xiong et al., 2026). AVLips is designed for audio-visual lip-sync forgery detection, with approximately 340,000 synthesized audio-visual samples. The dataset incorporates MakeItTalk (Zhou et al., 2020), Wav2Lip (Prajwal et al., 2020), TalkLip (Wang et al., 2023), and SadTalker (Zhang et al., 2023) to generate facial animations with realistic lip movements. Authentic content is collected from LRS3 (Afouras et al., 2018), FaceForensics++ (Rossler et al., 2019), DFDC (Dolhansky et al., 2020), and other real-world sources. Our experiments use the subset released by the authors, containing authentic LRS3 videos and GAN-generated forgeries. We divide this subset into training, validation, and test splits in a $6:1:3$ ratio. AATT, AVH-Align (Smeu et al., 2025), and T-AVFD (Liu et al., 2026) are trained on the same authentic videos from the training split. FKAV contains 500 authentic videos from VoxCeleb2 (Chung et al., 2018) and 19,500 manipulated videos. Visual manipulations are produced using FaceSwap (Korshunova et al., 2017), FSGAN (Nirkin et al., 2019), and Wav2Lip, while synthesized speech is generated using SV2TTS (Jia et al., 2018). The dataset therefore covers manipulations of either modality as well as joint audio-visual forgeries. Following the evaluation protocol of T-AVFD, we use a subset comprising 500 authentic and 1,000 manipulated videos spanning different generation methods. This subset reduces the pronounced class imbalance of the complete dataset. THB evaluates detection of generated portrait videos using FFHQ (Karras et al., 2019) images as source portraits and audio or video from CelebV-HQ (Zhu et al., 2022) as driving signals.

Its core collection covers six generation configurations, including Hallo (Xu et al., 2024a), Hallo2 (Cui et al., 2025), audio-driven AniPortrait, video-driven AniPortrait (Wei et al., 2024), LivePortrait (Guo et al., 2024), and EMOPortraits (Drobyshev et al., 2024). Additional generated samples are provided by MAGI-1 (Teng et al., 2025). The released data include separate training, validation, and test partitions for the respective generators. We combine the official test splits for all generation methods into a single evaluation set.

Singing Head. Singing Head DeepFake (SHDF) (Liu et al., 2026) extends detection evaluation to musical performances, where vocalization and facial motion differ from those in conversational speech. The complete dataset contains 2,600 authentic singing videos covering 80 identities and 3,000 synthesized videos covering 100 identities. Authentic performances are collected from YouTube, while additional portrait identities are drawn from HDTF. The synthesized videos are produced by MEMO, Hallo2, and EchoMimic, conditioned on recorded or synthesized singing audio. These generators animate facial expressions, lip movements, and head motion, providing diverse singing forgeries. We follow the evaluation protocol of T-AVFD and reserve SHDF for testing in the main comparison. For the training-distribution experiment, we instead train on authentic SHDF videos disjoint from the test set and evaluate the resulting model on both SHDF and AVLips.

A.2.4 Baseline Detection Methods

We compare AATT with four supervised detectors and three unsupervised detectors. The supervised methods require both authentic and manipulated training samples, whereas the unsupervised methods learn exclusively from authentic videos. We use released checkpoints for the supervised baselines and AVAD (Feng et al., 2023), and retrain AVH-Align (Smeu et al., 2025) and T-AVFD (Liu et al., 2026) on the same authentic training videos as AATT.

Supervised Baselines. Among the supervised baselines, CViT (Wodajo & Atnafu, 2021) combines convolutional feature extraction with a vision transformer to detect visual forgery cues. EfficientViT (Coccomini et al., 2022) similarly integrates an EfficientNet-B0 backbone with transformer-based classification, exploring different transformer architectures for visual deepfake detection. RealForensics (Haliassos et al., 2022) first learns representations from natural audio-visual correspondence in authentic videos, then uses these representations as auxiliary prediction targets during supervised real-fake classification. LipFD (Liu et al., 2024) instead uses both modalities to detect inconsistencies between speech and lip movements, incorporating relationships between lip and head regions to identify lip-sync forgeries.

Unsupervised Baselines. Among the unsupervised baselines, AVAD models sequences of audio-visual synchronization features autoregressively. Deviations from the temporal patterns learned on authentic videos provide evidence of manipulation. AVH-Align investigates the influence of leading silence in forged videos and learns audio-visual alignment from authentic pairs. It aggregates frame-level alignment scores into a video-level prediction, providing a baseline based on learned cross-modal correspondence. T-AVFD complements audio-visual alignment with semantic representations of facial authenticity. It learns these representations through alignment with textual descriptions and adaptively weights them alongside audio-visual features for detection.

A.3 Additional Experiments

A.3.1 Additional Ablation Studies

We further examine the training objective and transport direction. Table 7 compares three variants with the full model on SHDF and THB.

Training Objective. We replace the ranking loss with binary cross-entropy (BCE) (Mao et al., 2023), retaining the same matched audio-visual pairs and cross-video mismatches. This variant uses the same authentic training data and introduces no forged samples. BCE achieves results close to the full model on THB, but the gap widens on SHDF, where AUC decreases from 88.1% to 86.3%. The stronger performance of the ranking objective, particularly on singing videos unseen during training, supports directly optimizing the relative ordering of matched and mismatched pairs to learn transferable consistency evidence.

Table 7: **Additional ablation studies on SHDF and THB.** All variants are trained on the same authentic AVLips videos. We report AP (%) and AUC (%), with the best results in bold and the second-best results underlined.

Variant	SHDF *(singing)*		THB *(talking)*	
	AP	AUC	AP	AUC
BCE loss	92.1	86.3	<u>98.7</u>	<u>99.1</u>
Reverse transport	90.9	85.2	95.4	97.0
Bidirectional transport	<u>92.4</u>	<u>87.3</u>	97.5	98.2
AATT	**93.2**	**88.1**	**98.9**	**99.3**

Table 8: **Effect of the temporal window radius τ.** The radius is measured in feature time steps, with $\tau = 15$ used by default. We report AP (%) and AUC (%). Average denotes the unweighted mean across SHDF and THB. The best results are bold and the second-best results are underlined.

Radius τ	SHDF *(singing)*		THB *(talking)*		Average	
	AP	AUC	AP	AUC	AP	AUC
5	89.9	83.3	95.3	96.6	92.6	90.0
10	92.6	87.3	98.5	<u>98.9</u>	95.6	93.1
15 (default)	**93.2**	**88.1**	<u>98.9</u>	**99.3**	**96.1**	**93.7**
20	<u>93.0</u>	<u>87.7</u>	<u>98.9</u>	**99.3**	<u>96.0</u>	<u>93.5</u>
25	92.7	87.1	**99.0**	**99.3**	95.9	93.2

Transport Direction. The default formulation uses visual queries to aggregate audio features onto the visual timeline. Reverse transport instead uses audio queries to aggregate visual features, while bidirectional transport incorporates both directions. Reversing the transport direction reduces AP and AUC on both benchmarks, indicating that the two formulations are not interchangeable for detection. Bidirectional transport improves over the reverse variant but remains below the default formulation. These results support the adopted visual-query formulation, with no additional benefit from the evaluated bidirectional variant.

A.3.2 Effect of the Temporal Window Radius

The temporal window radius τ determines the range of audio candidates considered for each visual query. As shown in Table 8, increasing τ from 5 to 15 improves both AP and AUC on SHDF and THB. This trend suggests that a narrow candidate range limits the correspondence evidence available for consistency assessment.

Further increasing the radius beyond 15 offers little improvement on THB and reduces both AP and AUC on SHDF, indicating that a larger candidate range does not necessarily improve detection. The default setting, $\tau = 15$, achieves the strongest performance on SHDF while matching the highest AUC on THB, yielding the best average AP and AUC across the two benchmarks. These results support the adopted window radius for consistency assessment across talking and singing scenarios.

A.3.3 Discriminative Score Distributions

Figure 5 compares baseline detection scores with the descriptors C, S, and I and the joint detection score D obtained from the same full AATT model. The descriptors are evaluated directly without retraining, while D is produced by the learned scoring head. The behavior of mean correspondence informativeness I is particularly revealing. It achieves an AUC of 98.6% on talking videos but only 62.5% on singing videos, where the authentic and manipulated distributions overlap substantially. This contrast shows that the concentration of local matching preferences has different discriminative value across scenarios. It supports assessing how much information local evidence provides together with whether that evidence supports audio-visual consistency.

AATT addresses these complementary aspects by using correspondence informativeness r to regulate evidence contributions and combining I with transport compatibility C and path surprisal S.

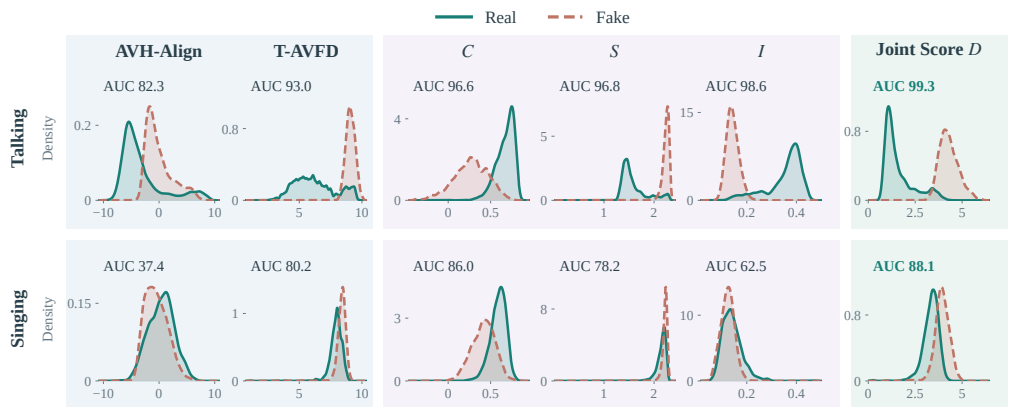

Figure 5: **Distributions of detection scores and correspondence descriptors.** Columns compare AVH-Align, T-AVFD, the individual AATT descriptors C, S, and I, and the joint detection score D. The upper and lower rows show talking (THB) and singing (SHDF) results. Solid teal and dashed coral curves represent authentic and manipulated videos. Each panel reports AUC (%), using $-C$, S, and $-I$ as manipulation scores for the descriptor panels. The curves display the original descriptor values.

The latter descriptors assess feature agreement after transport and support for temporally coherent correspondences, respectively. On singing videos, C retains stronger discrimination than either S or I, highlighting the value of examining transported feature agreement when correspondence informativeness alone is less discriminative. The joint score D achieves the highest AUC among both baselines and all individual descriptors in both scenarios. These results support integrating correspondence informativeness, temporal coherence, and transported feature agreement to obtain a more discriminative consistency measure across vocal settings.

A.4 Future Directions

The joint treatment of correspondence informativeness and temporal coherence in AATT suggests opportunities beyond deepfake detection. A natural extension is to adapt this formulation to audio-visual synchronization and cross-modal sequence retrieval. For synchronization, correspondence informativeness could guide the contribution of different temporal regions to alignment estimation. For retrieval, combining transported feature agreement with temporal structural evidence could provide a basis for comparing sequences with ambiguous local matches.

Another direction is to use the inferred correspondence posterior to guide multimodal representation learning, allowing training objectives to account for multiple plausible matches and their temporal dependencies. The descriptors developed in AATT also offer a basis for studying how correspondence concentration, feature agreement, and temporal support interact across tasks. Extending this perspective to other paired temporal modalities could broaden the applications of learning consistency from authentic multimodal observations.