MoCouple: Learning Fine-Grained Motion Coupling for Generalizable Talking-Face Forgery Detection

Anonymous authors
Paper under double-blind review

Abstract

Advances in talking face generation (TFG) enable increasingly convincing audio-visual impersonation, heightening the risks of identity misuse and privacy violations. Visual artifact analysis and audiovisual synchronization analysis constitute two dominant paradigms in current detection methods, with the latter primarily focusing on speech and lip dynamics. However, generator-specific artifacts can limit transfer to unseen synthesis methods, while accurate lip synchronization can render audio–lip mismatch less discriminative. Lip-centered analysis can also overlook subtle inconsistencies across facial regions and become vulnerable when non-frontal views obscure mouth dynamics. Although recent TFG methods increasingly capture audio–motion correspondences, aligning each motion component with audio alone does not ensure natural coordination across motions. Motivated by this, we propose **MoCouple** to detect subtle cross-component inconsistencies by learning the joint coordination of lip articulation, facial deformation, head movement, and speech from genuine videos. Specifically, masked latent reconstruction on real videos learns how lip articulation varies jointly with speech, facial deformation, and head motion, while window-level motion statistics capture weaker dependencies between audio and motion. Fusion guided by quality and evidence then combines these relation sensitive features with visual forensic cues to detect forgeries with plausible local motions but inconsistent joint dynamics. Experimental results demonstrate that, in addition to maintaining strong performance on the original training dataset, our method achieves significant out-of-distribution generalization capabilities that surpass existing methods. This work positions fine-grained motion coupling as a complementary authenticity criterion beyond visual fidelity and audio–lip alignment. The same perspective may inform the evaluation and training of TFG models, encouraging coherent joint dynamics rather than merely plausible movements of individual regions.

1 Introduction

Recent advances in talking face generation (TFG) have substantially lowered the barrier to synthesizing realistic videos of arbitrary identities from only a reference image. Compared with traditional face swapping, which manipulates existing facial regions, modern generators can directly synthesize complete facial appearance and motion. While these advances enable increasingly realistic and controllable generation, they also intensify the risks of identity impersonation, misinformation, and privacy violations, calling for detection paradigms that can adapt to the evolving characteristics of modern generative models (Zhang et al., 2023; Ji et al., 2025).

Existing TFG detectors mainly follow two paradigms. Visual forensic methods identify generation artifacts from facial appearance, local textures, or temporal discontinuities (Li et al., 2020; Zheng et al., 2021; Wang et al., 2023). Audio-visual methods detect inconsistencies between speech and facial motion, primarily through speech and lip synchronization or broader cross modal correspondence (Haliassos et al., 2021; Oorloff et al., 2024). However, visual artifacts are often specific to particular generation processes, which can limit generalization to unseen synthesis methods (Shiohara & Yamasaki, 2022). Meanwhile, modern generators explicitly optimize speech and lip synchronization and can produce increasingly accurate mouth dynamics (Prajwal et al., 2020; Zhou

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.

et al., 2021), weakening the discriminative power of local audiovisual inconsistencies. Moreover, a primary focus on lip dynamics can overlook subtle inconsistencies in the coordination among facial and head motions. Consequently, neither visual artifacts nor local audiovisual correspondence provides a sufficiently stable basis for detecting increasingly diverse TFG methods.

To seek a more generalizable detection cue, we analyze existing TFG models and identify a more fundamental challenge beyond visual artifacts and speech and lip synchronization, namely, reproducing the interactions among speech related motions during natural talking. Specifically, speech strongly constrains articulatory motion, while its relations with facial and head motions are statistically weaker. Meanwhile, these visual motion components also exhibit coordinated dependencies with each other (Yehia et al., 2002; Wang et al., 2021). These dependencies differ in strength and temporal structure, extending beyond simple frame level synchronization. Although modern generators can synthesize realistic individual motions, faithfully reproducing these joint dependencies remains challenging because identity, expression, pose, and speech driven motion are often modeled through partially disentangled generation processes (Zhou et al., 2021; Zhang et al., 2023; Peng et al., 2024). We therefore investigate the relationships among speech, vocal articulation, facial expression, and head motion, and find that their discriminative characteristics vary substantially between authentic and synthesized videos. This observation motivates us to move beyond conventional synchronization cues and model talking face authenticity through the joint dependencies among multiple motion components.

Based on this perspective, we propose **MoCouple**, a **Mo**tion-**Couple**d multimodal framework for generalizable TFG detection. MoCouple models fine-grained motion relationships through a dedicated relation stream and combines them with complementary visual forensic evidence. The two streams are further coupled according to their evidential reliability and conflict, allowing the detector to exploit complementary cues when individual evidence becomes weak or unreliable. Rather than asking only whether a talking face looks realistic or appears synchronized, MoCouple examines whether its underlying motions interact like those of natural talking behavior.

Our main contributions are as follows:

- We systematically analyze motion dependencies in real and generated talking faces and show that accurate local speech and lip synchronization does not necessarily imply natural coordination among multiple motion components.
- We introduce fine grained motion coupling as a new perspective for TFG detection, extending conventional speech and lip synchronization to the joint dependencies among speech, articulatory motion, facial deformation, and head motion.
- We propose **MoCouple**, which jointly models fine grained motion relationships and visual forensic evidence, and adaptively integrates them according to evidential reliability and conflict, leading to robust generalization across diverse forgery conditions.

2 Related Work

2.1 Talking Face Generation

Talking-face generation aims to synthesize realistic facial appearance and motion from a reference identity and speech audio. Early methods such as Wav2Lip (Prajwal et al., 2020) focused on accurate audio–lip synchronization, while later approaches introduced explicit 3D or neural representations for richer motion control. AD-NeRF (Guo et al., 2021) models talking heads with audio-conditioned neural radiance fields, SadTalker (Zhang et al., 2023) predicts 3D expression and head-pose coefficients, and SyncTalk (Peng et al., 2024) further coordinates lip articulation, expression, and head motion.

Recent diffusion-based methods such as EMO (Tian et al., 2024), Hallo (Xu et al., 2024a), VASA-1 (Xu et al., 2024b), and Hallo3 (Cui et al., 2025) further improve visual realism, expressiveness, and motion diversity. Many modern generators employ disentangled or factorized representations of identity, pose, expression, and speech-driven motion for controllable synthesis. As individual motion components and audio–lip synchronization become increasingly realistic, detection increasingly requires cues beyond coarse visual artifacts or local synchronization.

2.2 Visual Artifact-based Deepfake Detection

Visual forensic methods detect deepfakes from abnormal appearance and temporal traces. Earlier approaches primarily exploited blending, texture, and frequency artifacts, whereas recent methods increasingly adopt pretrained vision models such as CLIP (Radford et al., 2021) to improve generalization. DFD-FCG (Han et al., 2025), DFD-HR (Sun et al., 2026), and DiCoME (Kang et al., 2026) adapt CLIP representations through facial-component guidance, hierarchical feature routing, and semantic–artifact decomposition, respectively. Beyond pretrained semantic features, video-based methods exploit temporal generation traces: FakeSTormer (Nguyen et al., 2025) models spatial and temporal artifacts, while the training-free D3 (Zheng et al., 2025) detects generated videos through abnormal second-order temporal dynamics.

Although these methods improve robustness beyond conventional boundary artifacts, visual forensic cues can remain sensitive to generator-specific traces, compression, and visual quality. More fundamentally, they characterize how generated videos appear or evolve visually rather than whether different speech-related motions interact naturally.

2.3 Cross-Modal Audio-Visual Deepfake Detection

Audio-visual deepfake detectors exploit inconsistencies between speech and facial dynamics. Early multimodal methods jointly model auditory and visual streams and their correspondence (Zhou & Lim, 2021). Lips Are Lying (Liu et al., 2024) focuses on speech–lip temporal inconsistency, AVH-Align (Smeu et al., 2025) learns alignment between pretrained audio and visual representations, X-AVDT (Kim et al., 2026) exploits diffusion inversion and internal cross-attention cues, and LipDA (She et al., 2026) detects abnormal coupling between lip motion and head pose.

Despite this progress, existing approaches mainly focus on global audio–visual correspondence or a limited set of pairwise relationships. Lip-centric cues can become less discriminative as modern generators achieve increasingly accurate speech–lip synchronization, and may also degrade under low resolution, occlusion, large head poses, or inaccurate mouth localization (Haliassos et al., 2021). More broadly, the interactions among speech, vocal articulation, facial expression, and head motion remain comparatively underexplored, motivating our fine-grained motion-coupling formulation.

3 Fine-Grained Motion Coupling Analysis

Authentic talking-face behavior involves coordinated dynamics across speech and multiple facial motions (Chandrasekaran et al., 2009; Yehia et al., 2002; Munhall et al., 2004). Although modern generators achieve increasingly accurate local audio–lip synchronization, it remains unclear whether they preserve the broader motion dependencies of natural talking. We therefore study authentic and synthesized videos in learned fine-grained motion relation spaces.

3.1 Learning Authentic Motion Coupling

We characterize talking-face dynamics through three complementary relationships: Audio–Vocal, Audio–Motion, and Visual Motion. Audio–Vocal captures the strong correspondence between speech and vocal articulation, Audio–Motion describes weaker statistical dependencies between speech and broader facial dynamics, and Visual Motion characterizes coordination among vocal articulation, facial expression, and head motion. Together, these relations extend beyond conventional audio–lip synchronization toward a more fine-grained representation of natural talking behavior.

Motion Representation. Rather than using explicit geometric descriptors such as facial landmarks or 3D morphable model (3DMM) coefficients as the primary motion representation, we employ learned spatiotemporal features. Although explicit geometry provides interpretable measurements, its estimation can become unreliable under non-frontal poses, occlusion, poor illumination, or low visual quality, potentially introducing detection or fitting errors into subsequent relation modeling. Moreover, predefined geometric parameters provide a relatively constrained description of subtle facial dynamics. A qualitative analysis of these failure modes is provided in Appendix B.

We therefore employ BRAVEn (Haliassos et al., 2024), a self-supervised audio-visual speech representation model, to encode speech and vocal articulation, and VideoMAE (Tong et al., 2022), a masked video representation model, to capture broader spatiotemporal facial dynamics. Given the audio sequence $\mathbf{A}_{1:T}$, vocal-region sequence $\mathbf{V}_{1:T}$, and full-face sequence $\mathbf{F}_{1:T}$, the motion representations are

$$\begin{aligned} \mathbf{H}_a = E_a^{\mathrm{BRAVEn}}(\mathbf{A}_{1:T}), \quad \mathbf{H}_v = E_v^{\mathrm{BRAVEn}}(\mathbf{V}_{1:T}), \\ (\mathbf{H}_e, \mathbf{H}_h) = E_m^{\mathrm{VideoMAE}}(\mathbf{F}_{1:T}). \end{aligned} \tag{1}$$

Here, $\mathbf{H}_a$ and $\mathbf{H}_v$ encode acoustic and vocal-motion information, while $\mathbf{H}_e$ and $\mathbf{H}_h$ characterize expression- and head-motion-oriented dynamics. Two lightweight adapters specialize the shared acoustic features as $\mathbf{H}_a^v = P_v(\mathbf{H}_a)$ and $\mathbf{H}_a^m = P_m(\mathbf{H}_a)$, allowing speech information to support vocal articulation and broader facial motion differently. All relation-specific representations are projected to a common feature dimensionality before relation modeling. This dimensional unification facilitates heterogeneous relation modeling but does not impose explicit semantic alignment among the three relation spaces.

Authentic-Motion Pretraining. To learn natural motion dependencies without relying on generator-specific forgery artifacts, we pretrain the relation representation modules exclusively on authentic talking-face videos from LRS3 (Afouras et al., 2018) and HDTF (Zhang et al., 2021). Three groups of relation-oriented self-supervised objectives are employed. First, audio–vocal reconstruction predicts masked acoustic and vocal-motion representations from their available context. Second, visual-context reconstruction predicts masked vocal, expression, and head-motion representations from the remaining visual dynamics. Third, an audio–facial motion distribution objective predicts the mean and standard deviation of expression and head-motion features from acoustic motion representations, encouraging the underlying features to capture statistical speech–motion dependencies without imposing deterministic frame-wise correspondence.

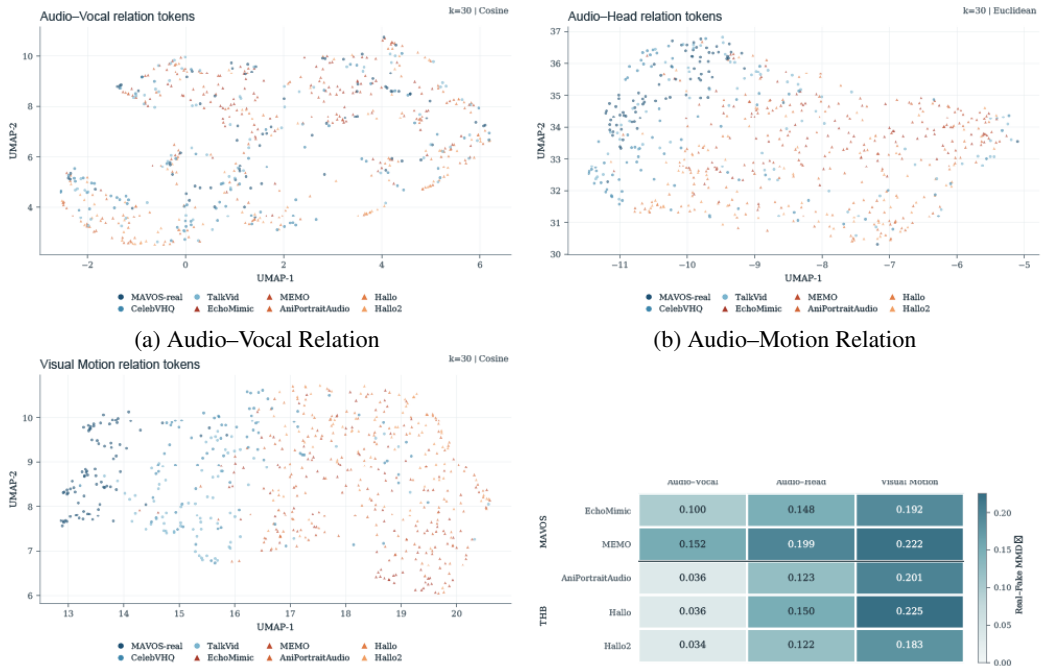
No real/fake classification supervision is introduced during this stage. The auxiliary objectives therefore establish representations of authentic speech-related motion dynamics before downstream forgery discrimination. The pretrained representation modules are frozen for the distribution analysis below and subsequently used to initialize the corresponding components of the Relation Stream in Sec. 4.3. Detailed masking strategies, auxiliary losses, and optimization settings are provided in Appendix C.2.

3.2 Motion Coupling Distribution Analysis

We freeze the real-only pretrained representation modules and extract video-level relation representations from the authentic and fully synthetic samples of Omni-Fake-Set AV-TH (Li et al., 2026). No Omni-Fake samples are used during the preceding representation learning stage. Window-level representations are aggregated at the video level and independently projected using UMAP (McInnes et al., 2018) for visualization. To avoid drawing conclusions solely from low-dimensional projections, we additionally quantify the discrepancy between the authentic reference distribution and each generator-specific synthetic distribution using unbiased RBF-kernel MMD^2 (Gretton et al., 2012) in the original feature spaces. Detailed computation and visualization settings are provided in Appendix A.

As shown in Fig. 1(a–c), authentic and synthesized samples substantially overlap in the Audio–Vocal space, whereas more apparent distributional differences emerge in the Audio–Motion and Visual Motion spaces. This observation suggests that contemporary talking-face generators can preserve relatively plausible audio–vocal correspondence while still deviating in weaker speech–facial-motion dependencies and higher-order visual motion coordination. The MMD^2 results in Fig. 1(d) further support the presence of representation shifts across different generation methods in the original feature spaces.

These observations motivate MoCouple to jointly model heterogeneous motion relationships rather than relying exclusively on audio–vocal correspondence. In the following section, the learned representations are further specialized for forgery discrimination and combined with complementary visual forensic evidence.



(a) Audio–Vocal Relation (b) Audio–Motion Relation

(c) Visual Motion Relation (d) Relation Distribution Discrepancy

Figure 1: **Motion-coupling distribution analysis on Omni-Fake-SET AV-TH.** (a–c) UMAP projections of Audio–Vocal, Audio–Motion, and Visual Motion representations extracted using the frozen representation modules pretrained exclusively on authentic videos. (d) MMD^2 between authentic and synthesized representations in the original feature spaces.

4 METHOD

4.1 PROBLEM DEFINITION

Given a talking-face video, we consider a temporal window of T frames and its corresponding audio segment. Let $\mathbf{A}_{1:T}$ denote the audio input, $\mathbf{F}_{1:T}$ the full-face sequence, and $\mathbf{V}_{1:T}$ the cropped vocal-region sequence. Region validity and quality information is additionally retained to account for unreliable visual observations. Our objective is to estimate the probability that the input window is forged:

$$\hat{y} = f_\theta\left(\mathbf{A}_{1:T}, \mathbf{F}_{1:T}, \mathbf{V}_{1:T}\right), \qquad \hat{y} \in [0, 1], \tag{2}$$

where $\hat{y}$ denotes the predicted fake probability. To this end, we jointly exploit fine-grained motion relationships and complementary visual forensic evidence.

4.2 OVERVIEW

As illustrated in Fig. 2, MoCouple consists of a Fine-Grained Relation Stream, a Visual Forensic Stream, and a Cross-Stream Multimodal Fusion module. The Relation Stream specializes the authentic-motion representations introduced in Sec. 3.1 for forgery discrimination, while the Forensic Stream captures complementary visual generation traces. Their outputs are subsequently integrated for final prediction.

The Relation Stream maps the audio, full-face, and vocal-region sequences to a Relation Token $\mathbf{z}_r$, while the Forensic Stream maps the two visual sequences to a Forensic Token $\mathbf{z}_f$. Each stream also produces a forgery prediction and estimates its class beliefs and uncertainty through an evidential head. Cross-stream fusion combines the two tokens with this evidential information for the final forgery prediction.

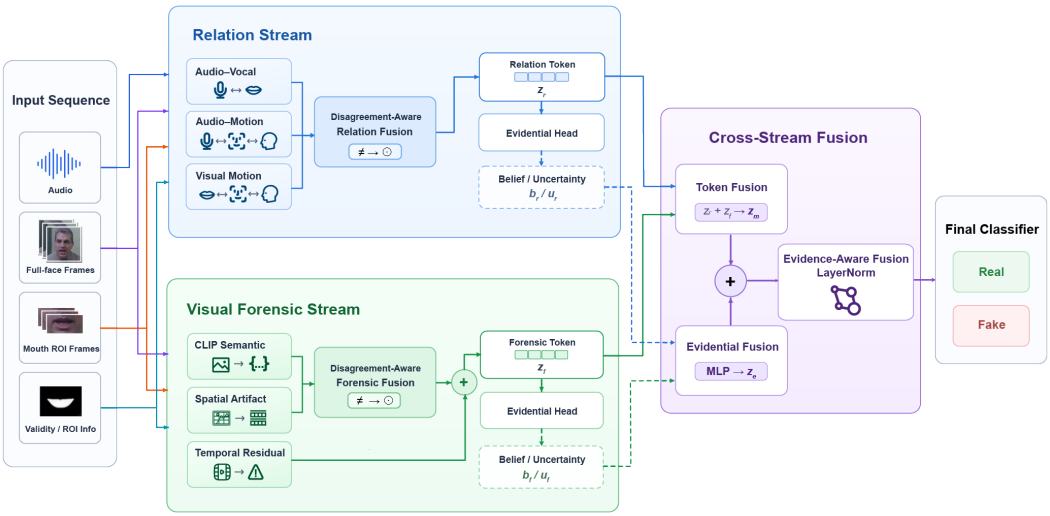


Figure 2: **Overview of MoCouple.** The Relation Stream specializes three pretrained motion representations for fine-grained relation modeling, while the Visual Forensic Stream combines semantic and spatial artifact evidence with temporal residual information. Transformer Feature Fusion integrates the stream representations, and Evidential Fusion incorporates their beliefs, uncertainties, and cross-stream conflict for forgery prediction.

4.3 Fine-Grained Relation Modeling

Fine-Grained Relation Experts. Using the motion representations introduced in Sec. 3.1, we construct three relation-specific Transformers for Audio–Vocal, Audio–Motion, and Visual Motion dependencies:

$$\begin{aligned} \mathbf{z}_{av} &= R_{av}(\mathbf{H}_a^v, \mathbf{H}_v), \\ \mathbf{z}_{am} &= R_{am}(\mathbf{H}_a^m, \mathbf{H}_e, \mathbf{H}_h), \\ \mathbf{z}_{vm} &= R_{vm}(\mathbf{H}_v, \mathbf{H}_e, \mathbf{H}_h). \end{aligned} \tag{3}$$

Unlike the real-only representation learning stage, these experts are optimized with forgery supervision to capture discriminative deviations from authentic motion relationships. Each expert produces a relation token and an auxiliary fake probability p_i, where $i \in \{av, am, vm\}$.

Disagreement-Aware Relation Fusion. Different forgery mechanisms may preserve some motion relationships while disturbing others, leading the three experts to produce conflicting predictions. We jointly model this disagreement and the reliability of each branch when forming the Relation Token:

$$\begin{aligned} \mathbf{d}_r &= [\mathrm{Var}(\mathbf{p}_r), \mathrm{range}(\mathbf{p}_r), |p_{av} - p_{am}|, |p_{av} - p_{vm}|, |p_{am} - p_{vm}|], \\ \mathbf{z}_r &= G_r\left(\mathbf{z}_{av}, \mathbf{z}_{am}, \mathbf{z}_{vm}, \mathbf{d}_r, \mathbf{q}_r, \mathbf{m}_r\right). \end{aligned} \tag{4}$$

Here, $\mathbf{p}_r = [p_{av}, p_{am}, p_{vm}]$, and $\mathrm{range}(\mathbf{p}_r) = \max(\mathbf{p}_r) - \min(\mathbf{p}_r)$. The descriptor $\mathbf{d}_r$ captures both overall prediction dispersion and pairwise differences, while $\mathbf{q}_r$ and $\mathbf{m}_r$ encode branch quality and availability. This allows the stream to account for heterogeneous and potentially unreliable relation evidence. A stream-level classifier and evidential head further produce the relation prediction, class beliefs, and uncertainty for subsequent cross-stream fusion. Details of branch-quality estimation and auxiliary supervision are provided in the supplementary material.

4.4 Visual Forensic Modeling

Visual Forensic Experts. We combine two complementary visual experts with an additional temporal residual pathway. A pretrained CLIP Vision Transformer (Radford et al., 2021) extracts multi-level semantic representations from the facial sequence, producing a semantic token $\mathbf{z}_c$. In parallel, a Spatial Artifact Expert extracts low-level generation traces from the full-face and vocal-region

sequences, producing an artifact token $\mathbf{z}_a$. The latter incorporates spatial, residual, and frequency-sensitive cues that may be suppressed in generic semantic representations. Each expert additionally produces an auxiliary fake probability, denoted by p_c and p_a. Detailed construction of the low-level artifact features is provided in the supplementary material.

Disagreement-Aware Forensic Fusion. Since semantic and artifact cues may respond differently to unseen generation methods, we condition their fusion on prediction disagreement and augment the result with temporal residual information:

$$\mathbf{z}_f = G_f\left(\mathbf{z}_c, \mathbf{z}_a, P_d(|p_c - p_a|)\right) + \mathbf{z}_t. \tag{5}$$

Here, P_d projects the absolute difference between the two expert predictions, and $\mathbf{z}_t$ encodes temporal residual features from the full-face and vocal-region sequences. The temporal residual pathway does not produce an independent prediction and therefore does not participate in the disagreement calculation. The resulting Forensic Token is further processed by a stream-level classifier and evidential head to obtain the forensic prediction, class beliefs, and uncertainty.

4.5 Cross-Stream Multimodal Fusion

The relation and forensic streams encode different aspects of talking-face authenticity and may exhibit different behavior across forgery conditions. We integrate them through feature-level interaction and an additional evidential representation derived from their stream-level predictions.

Transformer Feature Fusion. Given the Relation Token $\mathbf{z}_r$ and Forensic Token $\mathbf{z}_f$, we add learnable modality embeddings and introduce a CLS token before applying a Transformer encoder:

$$\mathbf{H}_m = F_{\mathrm{tr}}\left([\mathbf{z}_{\mathrm{cls}}, \mathbf{z}_r + \mathbf{e}_r, \mathbf{z}_f + \mathbf{e}_f]\right). \tag{6}$$

The encoded CLS representation captures cross-stream interactions, while a learned gate performs weighted aggregation over the modality tokens. Their combination yields the multimodal representation $\mathbf{z}_m$.

Evidential Fusion. Beyond feature representations, both streams produce Dirichlet-style evidential predictions (Sensoy et al., 2018), consisting of class beliefs $\mathbf{b}_s = [b_s^{\mathrm{real}}, b_s^{\mathrm{fake}}]$ and uncertainty u_s for $s \in \{r, f\}$. We augment the multimodal representation with these predictions and their cross-stream conflict:

$$\begin{aligned} c &= b_r^{\mathrm{real}} b_f^{\mathrm{fake}} + b_r^{\mathrm{fake}} b_f^{\mathrm{real}}, \\ \mathbf{z} &= \mathbf{z}_m + \mathrm{MLP}([\mathbf{b}_r, u_r, \mathbf{b}_f, u_f, c]). \end{aligned} \tag{7}$$

The conflict c measures the belief assigned to opposing classes across streams. Concatenating it with class beliefs and uncertainties yields a compact descriptor of stream-level confidence and conflict, whose projection complements feature-level fusion.

Training. Following the authentic-motion pretraining described in Sec. 3.1, the relation and forensic streams are optimized for forgery discrimination and subsequently integrated for multimodal detection. The overall objective includes final classification together with auxiliary stream- and expert-level supervision and representation regularization. Detailed evidential objectives, optimization stages, and hyperparameters are provided in the supplementary material.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate MoCouple on three talking-face forgery benchmarks: **MMDF** (Kim et al., 2026), **TalkingHeadBench (THB)** (Xiong et al., 2026), and **FakeAVCeleb** (Khalid et al., 2022). In addition, **HDTF** (Zhang et al., 2021) and **LRS3** (Afouras et al., 2018) are used exclusively for authentic-motion pretraining without forgery supervision. All supervised forgery training is conducted on MMDF. We follow the official splits for MMDF and THB and construct fixed evaluation subsets for FakeAVCeleb. The same evaluation partitions are used for all compared methods.

Table 1: Generalization across unseen generators and datasets in terms of video-level AUC (%).

		Cross-generator (MMDF)			Cross-dataset		Average†
	Method	**AniPortrait**	**MegActor-Σ**	**HunyuanAvatar**	**THB**	**FakeAVCeleb**	
Train-free	D3	69.67	41.79	31.92	47.20	49.54	47.98
	STALL	59.22	59.84	46.81	55.85	57.88	56.37
Visual	DFD-FCG	88.99	83.56	88.94	70.66	<u>95.56</u>	<u>84.47</u>
	DiCoME	98.89	**99.57**	**99.94**	<u>79.79</u>	71.93	83.73
Multimodal	X-AVDT*	<u>99.10</u>	90.17	<u>97.79</u>	24.21	**99.69**	73.06
	LipDA	63.68	86.13	78.28	53.48	55.93	61.80
	MoCouple	**99.36**	<u>99.36</u>	82.16	**95.76**	85.23	**91.54**

*MMDF and FakeAVCeleb results are taken from the original X-AVDT paper; THB is evaluated using the officially released checkpoint.

†The average is computed at dataset level, where MMDF generator results are aggregated into a single score.

Implementation Details. Each sample contains a 48-frame talking-face window with its corresponding 16-kHz audio segment, and all stream tokens are 128-dimensional. Detailed architecture configurations, optimization settings, sampling strategies, and training hyperparameters are provided in the supplementary material.

Evaluation Protocol. We use video-level area under the receiver operating characteristic curve (**AUC**) throughout the experiments. AUC avoids dependence on a fixed decision threshold and is therefore suitable for evaluating cross-domain generalization under potential score-calibration shifts. Window-level fake probabilities are averaged over all valid windows to obtain the final video-level score.

Baselines. We compare MoCouple with representative methods from three paradigms: training-free methods **D3** (Zheng et al., 2025) and **STALL** (Ben Hayun et al., 2026); visual methods **DFD-FCG** (Han et al., 2025), and **DiCoME** (Kang et al., 2026); and multimodal methods **X-AVDT** (Kim et al., 2026) and **LipDA** (She et al., 2026). Whenever available, we use the official implementations and follow the recommended configurations as closely as possible. For X-AVDT, whose official training setup is based on MMDF, we use the MMDF and FakeAVCeleb results reported in the original paper and evaluate additional datasets with the officially released checkpoint.

5.2 Generalization across Generators and Datasets

We evaluate generalization under two complementary settings using the same trained models without target-domain adaptation.

Cross-Generator Generalization. MMDF uses disjoint generation methods for training and testing. We therefore report video-level AUC separately for each generator in the test split to evaluate generalization to unseen synthesis methods.

Cross-Dataset Generalization. We further evaluate the same models directly on two external benchmarks (THB & FakeAVCeleb), introducing shifts in source videos, identities, recording conditions, and generation pipelines.

Results. Table 1 shows that MoCouple provides strong and balanced generalization across heterogeneous generation conditions. Although it does not achieve the highest AUC on every individual generator or dataset, its performance remains consistently strong across both cross-generator and cross-dataset settings, avoiding the pronounced degradation exhibited by several competing methods on particular target domains. MoCouple performs strongly on AniPortrait and MegActor-Σ and

Table 2: Ablation study of MoCouple in terms of video-level AUC (%) across three benchmarks.

Variant	Streams		Video-level AUC (%)		
	Relation	Forensic	MMDF	THB	FakeAVCeleb
Relation only	✓		89.60	**97.09**	83.98
Forensic only		✓	91.05	85.80	78.20
Full MoCouple	✓	✓	**93.63**	95.76	**85.23**

remains effective on HunyuanAvatar, while cross-dataset evaluation shows particularly robust transfer to THB and retaining solid performance on FakeAVCeleb. This balanced behavior results in the highest dataset-level average AUC of 91.54%, highlighting the robustness of MoCouple across different generators and datasets rather than specialization to a particular target domain. A more challenging MMDF-to-AVLips (Liu et al., 2024) cross-domain setting is investigated in Appendix D, including further analyses of the substantial visual and temporal distribution shifts observed in this setting.

5.3 Ablation Studies

We evaluate the contributions of the Relation and Forensic Streams by comparing each individual stream with the full MoCouple under the same data splits and evaluation protocol, as shown in Table 2.

The two streams exhibit distinct transfer characteristics. The Relation Stream performs particularly strongly on the two external benchmarks, suggesting that fine-grained motion dependencies provide relatively stable cues across dataset shifts. In contrast, the Forensic Stream achieves stronger performance on MMDF but degrades more noticeably on THB and FakeAVCeleb, indicating greater sensitivity to dataset-specific visual statistics.

Combining the two streams improves performance on MMDF and FakeAVCeleb, while yielding a small decrease relative to the Relation Stream alone on THB. This suggests that visual forensic evidence provides useful complementary information but is more domain-dependent, whereas relation-based motion cues constitute the more stable cross-domain component. Overall, the full model provides a more balanced trade-off across the evaluated domains.

6 Conclusion

We presented MoCouple, a generalizable talking-face forgery detector that examines authenticity through fine-grained motion coupling. By modeling the relationships among speech, vocal articulation, facial expression, and head motion, our analysis shows that different motion dependencies exhibit substantially different discriminative characteristics, with broader speech–facial and visual-motion relationships revealing more pronounced distributional deviations between authentic and synthesized videos.

Motivated by this observation, MoCouple combines a Fine-Grained Relation Stream with complementary visual forensic evidence through cross-stream multimodal fusion. Experiments across unseen generators and datasets demonstrate balanced and relatively stable generalization across heterogeneous target domains. Ablation results further show that relation-based motion cues provide the more stable cross-domain evidence, while visual forensic cues offer useful but more domain-dependent complementary information. Overall, our results highlight the value of examining how different components of talking-face behavior interact, rather than assessing their plausibility in isolation.

AI use statement

In this work, we used generative AI tools to assist with refining research hypotheses and experimental designs, discussing methodological choices, implementing and debugging research code, and

interpreting experimental results. We additionally used generative AI tools for literature search and summarization, manuscript drafting and language editing, and improving the presentation of figures, tables, and mathematical descriptions. We did not use generative AI tools to autonomously conduct experiments, fabricate experimental results, or generate evaluation data reported in this work.

All AI-assisted outputs were reviewed and verified by the authors. Suggested methodological ideas were assessed against relevant literature and validated through empirical experiments. AI-assisted code was manually inspected, executed, and tested before being used in experiments. References and factual claims suggested during literature search were checked against the corresponding original sources. All reported quantitative results were obtained from experiments conducted by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

Ethics statement

Talking-face generation technologies have legitimate applications in digital content creation and human–computer interaction, but they can also facilitate identity impersonation, misinformation, and other forms of misuse. This work studies detection methods for identifying such manipulated content and is intended to support media authenticity and forensic analysis. Our experiments are conducted using existing research datasets and generated talking-face data for research purposes, without collecting new personally identifiable data from human subjects. We do not release tools intended to facilitate the creation or malicious use of forged media. Nevertheless, as with other forgery detection systems, the proposed method may exhibit performance variations across identities, recording conditions, datasets, and generation methods. We therefore caution against treating its predictions as definitive evidence in high-stakes forensic or legal applications without additional human and contextual verification.

Reproducibility statement

We provide detailed descriptions of the proposed architecture, motion-coupling formulation, fusion mechanisms, and training objectives in the main paper. The experimental setup, datasets, evaluation protocols, and metrics are described in the experimental section. Additional implementation details, including data preprocessing, model configurations, training settings, and evaluation procedures, are provided in the appendix to facilitate reproduction of our results. We plan to organize and publicly release the implementation on GitHub after publication.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition, 2018. URL https://arxiv.org/abs/1809.00496.

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free detection of generated videos via spatial-temporal likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16299–16310, June 2026.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Chandramouli Chandrasekaran, Andrea Trubanova, Sébastien Stillittano, Alice Caplier, and Asif A Ghazanfar. The natural statistics of audiovisual speech. *PLoS computational biology*, 5(7): e1000436, 2009.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21086–21095, June 2025.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. Towards fast, accurate and stable 3d dense face alignment. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 152–168, Cham, 2020. Springer International Publishing.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5764–5774. IEEE, 2021.

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, Andreas Zinonos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Braven: Improving self-supervised pre-training for visual and auditory speech recognition, 2024. URL https://arxiv.org/abs/2404.02098.

Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22995–23005, 2025. doi: 10.1109/CVPR52734.2025.02141.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 193–203. IEEE, 2025.

Xiaolu Kang, Zhongyuan Wang, Jikang Cheng, Baojin Huang, Zhanhe Lei, Gang Wu, Qin Zou, and Qian Wang. Divide and conquer: Reliable multi-view evidential learning for deepfake detection, 2026. URL https://arxiv.org/abs/2606.01885.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus, 2019. URL https://arxiv.org/abs/1907.06724.

Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022. URL https://arxiv.org/abs/2108.05080.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

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 (CVPR)*, pp. 4403–4414, June 2026.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5000–5009. IEEE, 2020.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Xinze Li, Bingyu Zhu, Wuhui Duan, Congang Chen, Zeyu Fu, Yi Dong, Baoyuan Wu, Jason Li, and Guangliang Cheng. Omnifake: Benchmarking unified multimodal social media deepfake detection, 2026. URL https://arxiv.org/abs/2605.01638.

Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, Ziyou Liang, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91131–91155. Curran Associates, Inc., 2024. doi: 10.52202/079017-2892. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a5a5b0ff87c59172a13342d428b1e033-Paper-Conference.pdf.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Kevin G Munhall, Jeffery A Jones, Daniel E Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science*, 15(2):133–137, 2004.

Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10786–10796, October 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27092–27102. IEEE, 2024.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 666–676. IEEE, 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a981f2b708044d6fb4a71a1463242520-Paper.pdf.

Tianyi She, Jiawei Liu, Weifeng Liu, Hanqing Zhao, Weiming Zhang, and Kejiang Chen. Ariadne's thread of lipsync: Unraveling forgeries via inconsistency between lip motions and head poses. In *Forty-third International Conference on Machine Learning*, 2026.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, June 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18815–18825, June 2025.

Jiamu Sun, Zhiyuan Yan, Ke-Yue Zhang, Taiping Yao, and Shouhong Ding. Dfd-hr: Generalizable deepfake detection via hierarchical routing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13984–13995, June 2026.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive – generating expressive portrait videos with audio2video diffusion model under weak conditions, 2024. URL https://arxiv.org/abs/2402.17485.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. doi: 10.52202/068431-0732. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/416f9cb3276121c42eebb86352a4354a-Paper-Conference.pdf.

Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4129–4138. IEEE, 2023.

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, 2026. doi: 10.1109/WACV61042.2026.00403.

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, 2024a. URL https://arxiv.org/abs/2406.08801.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 660–684. Curran Associates, Inc., 2024b. doi: 10.52202/079017-0021. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/014fe398da515cd552fa6e1f33e0565e-Paper-Conference.pdf.

Hani C Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of phonetics*, 30(3):555–568, 2002.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8652–8661. IEEE, 2023.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12852–12862, October 2025.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15024–15034. IEEE, 2021.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4174–4184. IEEE, 2021.

Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14800–14809, October 2021.

A Details of Motion Coupling Distribution Analysis

This section provides additional implementation details for the motion-coupling distribution analysis presented in the main paper. The analysis is designed to investigate whether different talking-face generation methods preserve the motion relationships naturally observed in authentic videos, rather than to directly evaluate the classification capability of MoCouple.

A.1 Relation Representation Extraction

The analysis uses the Relation Stream obtained after real-only self-supervised pre-training. The pre-training data consist exclusively of authentic talking-face videos from LRS3 and HDTF, and no forgery-classification objective is introduced at this stage. All model parameters are frozen during the subsequent distribution analysis.

For each temporal window, we extract representations corresponding to the three fine-grained relation spaces considered in MoCouple: Audio–Vocal, Audio–Motion, and Visual Motion. These relation spaces capture complementary dependencies between speech and vocal articulation, speech and broader facial motion, and different visual motion components, respectively.

Because videos generally contain multiple overlapping windows, the window-level relation representations belonging to the same video are aggregated into a single video-level representation before distribution analysis. Consequently, each point shown in the visualization corresponds to one video rather than one temporal window, preventing long videos from disproportionately dominating the distribution.

A.2 UMAP Visualization

We use Uniform Manifold Approximation and Projection (UMAP) to qualitatively visualize the distributions of authentic and synthesized relation representations. UMAP is independently fitted for each relation space using all video-level representations included in the corresponding analysis.

Neither real/fake labels nor generator identities are provided to UMAP during fitting. Therefore, the resulting separation is induced solely by the geometry of the learned representation space rather than by supervised information introduced during dimensionality reduction.

We use a fixed random seed for reproducibility. The Audio–Vocal representation is visualized using cosine distance with 30 neighbors, the Audio–Motion representation using Euclidean distance with 30 neighbors, and the Visual Motion representation using cosine distance with 30 neighbors.

The two-dimensional UMAP projections are used only for qualitative visualization. All quantitative distribution comparisons are performed directly in the original high-dimensional relation spaces.

A.3 Distribution Discrepancy

To quantify the discrepancy between authentic and synthesized relation representations, we use the squared Maximum Mean Discrepancy (MMD^2) with an RBF kernel.

Given authentic representations $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{m}$ and synthesized representations $\mathcal{Y} = \{\mathbf{y}_j\}_{j=1}^{n}$, we compute the unbiased estimator

$$\begin{aligned}\widehat{\mathrm{MMD}}_u^2(\mathcal{X},\mathcal{Y}) = &\frac{1}{m(m-1)}\sum_{i\neq i'} k(\mathbf{x}_i,\mathbf{x}_{i'}) \\ &+ \frac{1}{n(n-1)}\sum_{j\neq j'} k(\mathbf{y}_j,\mathbf{y}_{j'}) \\ &- \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(\mathbf{x}_i,\mathbf{y}_j),\end{aligned} \tag{8}$$

where the RBF kernel is defined as

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\sigma^2}\right). \tag{9}$$

For the quantitative analysis reported in the main paper, authentic videos are pooled into a common reference distribution. MMD^2 is then independently calculated between the authentic reference distribution and the synthesized distribution produced by each generation method.

A larger MMD^2 indicates a stronger distributional discrepancy between authentic and synthesized motion-coupling representations. Since MMD is computed in the original representation space, the quantitative conclusions do not depend on the particular two-dimensional UMAP projection shown for visualization.

A.4 Statistical Testing

To assess whether the observed distribution discrepancies can be explained by finite-sample variation, we additionally perform permutation testing on the MMD^2 statistic. For each authentic-versus-synthesized comparison, the authentic and synthesized samples are first pooled, after which their group assignments are randomly permuted while preserving the original group sizes. MMD^2 is then recomputed under each permutation to construct an empirical null distribution.

We use 500 permutations for each comparison. The empirical p-value is determined by comparing the observed MMD^2 with the statistics obtained under random reassignment. Since multiple generators and relation spaces are evaluated, we apply the Benjamini–Hochberg procedure to control for multiple comparisons(Benjamini & Hochberg, 1995).

The permutation tests are used only as supplementary quantitative evidence for the distribution discrepancies observed in the learned relation spaces. They do not affect representation extraction, UMAP fitting, sample selection, or the configurations used for visualization. In particular, all statistical comparisons are performed using the original high-dimensional representations rather than their two-dimensional UMAP projections.

A.5 Robustness to UMAP Configurations

Because UMAP visualization can be affected by its neighborhood size and distance metric, we additionally examine several configurations to verify that the qualitative observations in the main paper are not specific to a single projection setting. For each of the three relation spaces, we consider neighborhood sizes $n_{\text{neighbors}} \in \{15, 30\}$ together with cosine and Euclidean distance, resulting in four candidate configurations for each relation representation.

Across these configurations, the principal qualitative structure remains consistent. Audio–Vocal representations generally exhibit substantial overlap between authentic and synthesized samples, whereas the distribution shifts are more visually apparent in the Audio–Motion and Visual Motion relation spaces. Changing the neighborhood size mainly affects the balance between local cluster compactness and global organization, while changing the distance metric alters the relative geometry of the projected clusters without changing the overall observation.

For the figures reported in the main paper, we use cosine distance with 30 neighbors for Audio–Vocal, Euclidean distance with 30 neighbors for Audio–Motion, and cosine distance with 30 neighbors for Visual Motion. Among the examined configurations, these settings provide clearer visualization of the local neighborhood structure and overall distribution geometry.

We emphasize that UMAP serves only as a qualitative visualization tool in our analysis. Consequently, the exact appearance of individual clusters is not used as quantitative evidence. All reported distribution discrepancies and statistical tests are computed in the original high-dimensional relation spaces using MMD^2, making the quantitative conclusions independent of the particular UMAP configuration used for presentation.

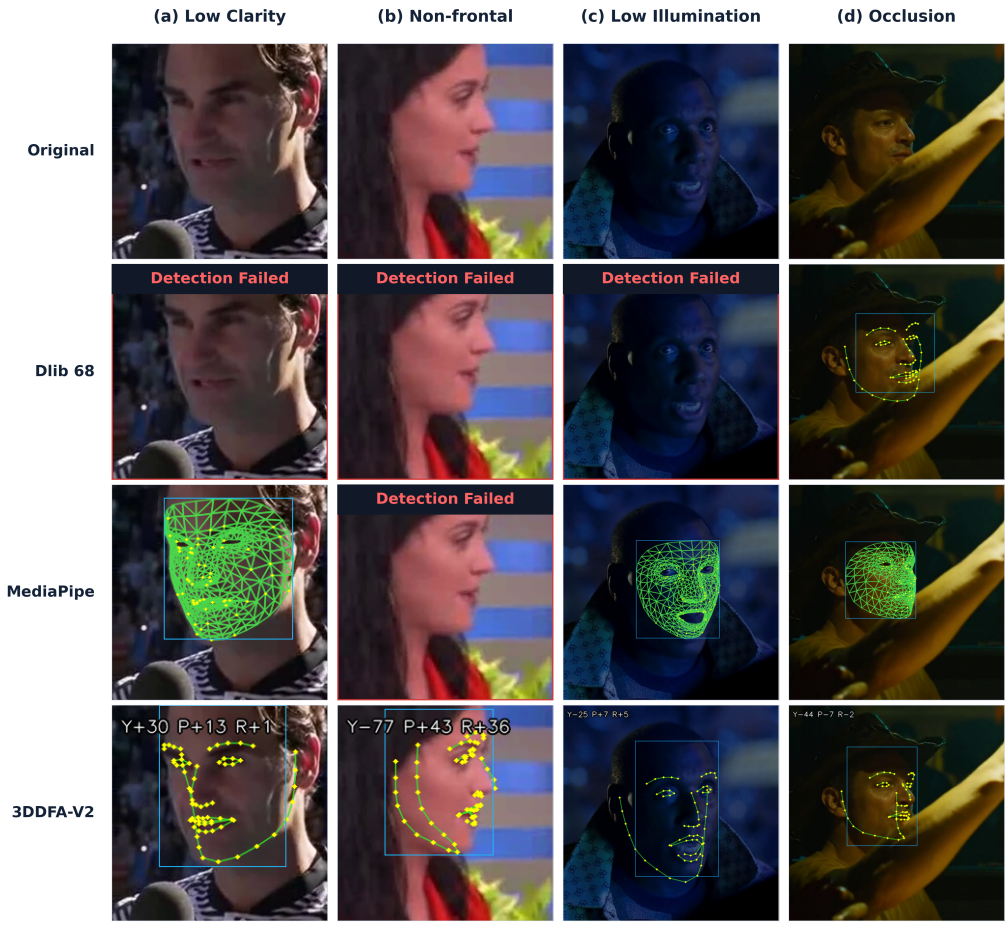


Figure 3: Qualitative reliability analysis of explicit facial geometry estimation on challenging authentic talking-face videos. We compare Dlib 68-point landmarks, MediaPipe Face Mesh, and 3DDFA-V2 under low clarity, non-frontal pose, low illumination, and occlusion. "Detection Failed" indicates that the corresponding method does not return a valid facial geometry estimate. Even when detection succeeds, noticeable localization errors may remain, particularly around the facial contour under difficult conditions.

B Reliability Analysis of Explicit Facial Geometry

Explicit facial landmarks and 3D face fitting provide intuitive geometric representations of facial motion, and are therefore natural candidates for modeling lip articulation, facial expression, and head movement. However, their reliability can degrade substantially under unconstrained visual conditions. To examine this issue, we conduct a qualitative case study using only authentic talking-face videos, thereby avoiding potential confounding effects from manipulation-specific artifacts.

We consider four representative challenging conditions: *low clarity*, *non-frontal pose*, *low illumination*, and *occlusion*. For each case, we compare three representative facial geometry estimators with different technical characteristics: Dlib 68-point landmarks(Kazemi & Sullivan, 2014), MediaPipe Face Mesh(Kartynnik et al., 2019), and 3DDFA-V2(Guo et al., 2020). Figure 3 shows representative results.

B.1 Qualitative Case Analysis

Dlib 68-point landmarks. Among the examined cases, Dlib exhibits the weakest robustness to challenging visual conditions. Detection failure rate is relatively high under low clarity, large non-frontal pose, and low illumination. Occlusion is comparatively less severe in the example shown in Fig. 3(d), where the detector still produces landmarks; however, the estimated geometry is visibly displaced from the actual facial structure. We note that complete Dlib detection failures are also observed for other occluded frames, while the successful example is retained here to illustrate that successful detection does not necessarily imply accurate localization.

MediaPipe Face Mesh. MediaPipe is generally the most stable and visually accurate method among the three estimators in the selected examples. It successfully reconstructs dense facial geometry under low clarity and remains relatively robust under low illumination and partial occlusion. Nevertheless, its robustness is not unconditional. A sufficiently large non-frontal pose leads to complete detection failure, while low illumination and occlusion can still introduce visible geometric deviations. These results suggest that dense landmark estimation improves robustness but does not eliminate failure under severe appearance or visibility changes.

3DDFA-V2. 3DDFA-V2 achieves the highest detection coverage among the selected challenging cases and successfully returns a fitted face for all examples shown in Fig. 3. However, successful fitting does not necessarily correspond to accurate geometric recovery. Noticeable fitting errors are present across several difficult cases. In particular, the estimated facial contour can deviate substantially from the true boundary, with contour landmarks occasionally projected toward the interior of the face. Similar shifts can also be observed around local facial structures under non-frontal pose, low illumination, and occlusion. This behavior is particularly relevant for motion modeling, since a continuously available but systematically misaligned geometry may introduce misleading motion trajectories.

B.2 Discussion

This qualitative analysis illustrates the vulnerability of explicit facial geometry estimation under challenging talking-face conditions. The three methods expose two distinct types of failure. The first is *detection failure*, where no valid facial landmarks or geometry are produced. This behavior is most evident for Dlib and, under large non-frontal poses, MediaPipe. The second is *geometric estimation error*, where a method successfully returns an output but the estimated geometry is visibly inconsistent with the actual facial structure. 3DDFA-V2 is particularly illustrative of the latter case: its high detection coverage avoids missing observations, but difficult inputs can still produce inaccurate facial contours and local landmark positions.

For motion analysis, both failure modes are problematic. Detection failure creates missing observations in the temporal trajectory, whereas inaccurate fitting can introduce artificial displacement that may be incorrectly interpreted as genuine lip, expression, or head motion. The latter can be especially difficult to identify automatically because the estimator still produces a formally valid output.

These observations motivate our decision not to use explicit facial landmarks or fitted 3D geometry as a mandatory intermediate representation in MoCouple. Instead, vocal, expression, and head-motion representations are derived directly from learned visual features. This design preserves richer appearance and temporal context while avoiding a hard dependency on the success and geometric accuracy of an external landmark or 3D fitting pipeline.

C Training Details

MoCouple is optimized in a stage-wise manner. The Relation and Forensic streams are first trained independently according to their respective objectives, after which their representations are combined by the cross-stream fusion module.

The complete training procedure consists of four stages: (1) self-supervised pre-training of the Relation Stream on authentic talking-face videos, (2) supervised Relation Stream training for forgery

detection, (3) independent supervised training of the Forensic Stream, and (4) cross-stream fusion with both pretrained streams frozen.

This staged design allows each stream to establish its own representation before multimodal fusion. In particular, the Relation Stream first learns normal motion-coupling patterns from authentic data, while the Forensic Stream independently learns complementary visual forensic cues. The final fusion stage is then responsible for combining the two sources of evidence rather than learning all representations jointly from scratch.

C.1 Input Sampling and Preprocessing

All training, validation, and evaluation stages operate on fixed-length temporal windows of $T = 48$ frames with a stride of 12 frames.

For each temporal window, the corresponding audio segment is determined according to the timestamp of the first video frame. Audio is represented as a 16 kHz mono waveform. For a nominal frame rate of 25 fps, a 48-frame window corresponds to approximately 1.92 seconds and therefore contains 30,720 audio samples. The actual audio segment is nevertheless aligned according to the video timestamps rather than assuming that every original video has an identical frame rate.

Full-face images are resized to 224×224. Mouth regions are dynamically cropped according to the detected mouth boxes. The Relation Stream uses 88×88 mouth crops for motion representation, whereas the Forensic Stream uses 128×128 crops to preserve additional local appearance information for forensic analysis.

A temporal window is considered to contain sufficiently reliable mouth information when it contains at least eight valid mouth observations and at least four consecutively valid frames.

When an audio track is unavailable, a zero waveform is used together with an explicit audio-availability indicator. The model directly consumes waveform and visual observations; no precomputed MFCC features, head-pose trajectories, or facial-expression trajectories are required. Expression and head-motion representations are instead derived from the visual encoder.

C.2 Relation Stream Pre-training

The Relation Stream is first pretrained using only authentic talking-face videos from LRS3 and HDTF. The objective of this stage is to establish a representation of normal motion coupling before introducing any forgery-classification supervision.

The audio and vocal encoders are initialized from the corresponding BRAVEn Base modality checkpoints, while the facial-motion encoder is initialized from VideoMAE Base. Given the encoded audio, vocal, expression, and head-motion representations, we construct self-supervised objectives that model their natural dependencies.

Masked latent reconstruction. Valid latent positions are randomly masked with probability 0.3, after which the model reconstructs the missing representation using information from the corresponding coupled context.

The reconstruction objectives contain two complementary groups. First, Audio–Vocal reconstruction models the strong relationship between speech content and vocal articulation. Second, visual-motion reconstruction models dependencies among vocal motion, facial expression, and head motion. These objectives encourage the Relation Stream to model coordinated motion structure rather than isolated modality-specific features.

Window-level motion statistics prediction. In addition to masked reconstruction, the audio-motion representation is used to predict summary statistics of facial expression and head motion over the temporal window. Specifically, auxiliary predictors estimate the mean and standard deviation of the corresponding visual-motion latents.

Both reconstruction targets and motion-statistics targets are detached from the computational graph, encouraging contextual prediction rather than allowing the prediction and target representations to trivially co-adapt.

The overall pre-training objective is

$$\mathcal{L}_{\text{pre}} = \sum_i \lambda_i \mathcal{L}^i_{\text{rec}} + \lambda_{\text{stat}} \mathcal{L}_{\text{stat}}, \tag{10}$$

where $\mathcal{L}^i_{\text{rec}}$ denotes the active masked latent reconstruction objectives and $\mathcal{L}_{\text{stat}}$ denotes window-level motion-statistics prediction.

The internal weights of the six auxiliary terms are 0.5, 1.0, 1.0, 0.6, 0.2, and 0.2, respectively. Active terms are combined by weighted averaging.

No forgery-classification objective is applied during Relation pre-training. Components that are not directly connected to these self-supervised objectives, including the final classification and relation-level fusion components, are mainly learned during the subsequent supervised stage. We therefore regard this stage as learning a normal motion-coupling prior rather than pretraining the complete Relation-based forgery detector.

C.3 Supervised Relation Stream Training

The Relation Stream is initialized from the self-supervised checkpoint and subsequently optimized on authentic and synthesized talking-face videos for forgery classification.

Both the individual fine-grained relation branches and their fused representation receive supervision. This encourages each type of motion relationship to retain discriminative information while allowing the final Relation representation to combine complementary relation cues.

The supervised Relation objective is

$$\begin{aligned}\mathcal{L}_{\text{rel}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{branch}} \mathcal{L}_{\text{branch}} + \lambda_{\text{sup}} \mathcal{L}_{\text{supcon}} \\ + \lambda_{\text{edl}} \mathcal{L}_{\text{edl}} + \lambda_{\text{aux}} \mathcal{L}^{\text{real}}_{\text{pre}},\end{aligned} \tag{11}$$

where $\mathcal{L}_{\text{cls}}$ denotes binary cross-entropy applied to the fused Relation prediction, $\mathcal{L}_{\text{branch}}$ supervises the individual fine-grained relation branches, $\mathcal{L}_{\text{supcon}}$ is a supervised contrastive objective(Khosla et al., 2020), and $\mathcal{L}_{\text{edl}}$ provides evidential supervision.

The self-supervised pre-training objective is retained with a small coefficient during supervised training, but is evaluated only on authentic samples. This encourages the Relation Stream to preserve the normal coupling structure learned during pre-training while adapting to forgery discrimination, without explicitly forcing synthesized samples to satisfy the authentic motion-coupling prior.

Evidential optimization. The evidential objective includes a KL regularization term whose coefficient is linearly increased during the first five epochs until reaching 0.1. This gradual schedule avoids imposing strong evidential regularization before the classification representation becomes sufficiently stable.

Optimization strategy. We adopt gradual unfreezing for the pretrained modality encoders. The BRAVEn audio and vocal backbones are frozen for the first three epochs and are subsequently optimized with a learning rate of 1×10^{-5}. VideoMAE is also frozen during the first three epochs; afterwards, only its final three Transformer blocks and normalization layers are unfrozen and optimized with a learning rate of 5×10^{-6}. The remaining Relation modules are optimized with a learning rate of 1×10^{-4}.

This asymmetric learning-rate strategy allows newly introduced relation modules to adapt relatively quickly while limiting destructive updates to pretrained modality representations.

C.4 Forensic Stream Training

The Forensic Stream is trained independently from the Relation Stream. Its objective is to capture visual generation evidence that is complementary to motion-coupling inconsistencies.

The stream contains two principal experts: a CLIP-based semantic expert and a spatial artifact expert. Temporal forensic information is introduced through an additional residual pathway rather than being treated as a third parallel expert.

Semantic expert. The semantic expert uses CLIP ViT-B/16 as its visual backbone. For each temporal window, eight frames are sampled for CLIP feature extraction. Intermediate patch-level representations are retained rather than relying only on the final global CLIP representation, preserving local semantic information around facial regions while still benefiting from the high-level representation learned by CLIP.

The pretrained CLIP backbone remains frozen throughout Forensic training, while newly introduced projection and adaptation components are optimized.

Spatial artifact expert. The spatial artifact expert analyzes both the full face and the local vocal region to capture generation artifacts that may not be sufficiently represented by high-level semantic features. The semantic and spatial artifact representations are subsequently combined by the Forensic fusion module.

Temporal artifact residual. Temporal forensic cues are extracted from frame-to-frame changes, high-frequency variations, and temporal-frequency characteristics. Instead of producing a third independent expert prediction, the temporal artifact feature is injected into the fused Forensic representation through a learnable residual coefficient. This keeps the primary Forensic structure centered on semantic and spatial evidence while allowing temporal inconsistency to refine the final visual token.

The complete Forensic objective is

$$\begin{aligned}\mathcal{L}_{\text{for}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{exp}}\mathcal{L}_{\text{expert}} + \lambda_{\text{sup}}\mathcal{L}_{\text{supcon}} \\ + \lambda_{\text{edl}}\mathcal{L}_{\text{edl}} + \lambda_{\text{decorr}}\mathcal{L}_{\text{decorr}},\end{aligned} \tag{12}$$

where $\mathcal{L}_{\text{cls}}$ supervises the final Forensic prediction, $\mathcal{L}_{\text{expert}}$ provides classification supervision for the semantic and spatial artifact experts, $\mathcal{L}_{\text{supcon}}$ regularizes the learned Forensic embedding, and $\mathcal{L}_{\text{edl}}$ provides evidential supervision.

We additionally employ $\mathcal{L}_{\text{decorr}}$ to discourage excessive redundancy between the semantic and spatial artifact tokens. In implementation, this term penalizes their squared cosine similarity.

Optimization strategy. The CLIP ViT-B/16 backbone remains frozen throughout the entire stage. Trainable components include the feature-projection layers, spatial artifact encoders, temporal artifact modules, forensic fusion components, and prediction heads. These modules are optimized with a learning rate of 3×10^{-4}.

C.5 Cross-Stream Fusion Training

After the Relation and Forensic streams have been trained independently, both streams are frozen and the cross-stream fusion module is optimized separately.

Each stream outputs a 128-dimensional representation. The resulting Relation and Forensic tokens are processed by a two-layer Transformer encoder with four attention heads, allowing information exchange between motion-coupling and visual-forensic evidence.

Freezing the two streams during this stage prevents their independently learned representations from being immediately overwritten by the fusion objective, allowing the optimization to focus on learning how the two sources of evidence should be combined.

Evidence-aware fusion. MoCouple additionally incorporates evidential information generated by the individual streams. Each stream predicts belief and uncertainty information. Together with cross-stream conflict measurements, these statistics form a seven-dimensional evidence representation. The evidence metadata are projected into the token embedding space and incorporated into the fusion process.

This allows the fusion module to consider not only the content of the Relation and Forensic representations, but also their reliability and degree of agreement.

Optimization objective. The Fusion loss is

$$\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{cls}} + 0.10\mathcal{L}_{\text{supcon}} + 0.05\mathcal{L}_{\text{stream}} + 0.05\mathcal{L}_{\text{cons}}, \tag{13}$$

Table 3: Loss weights used in the supervised training stages of MoCouple.

Stage	Loss Term	Weight
Relation	$\mathcal{L}_{\text{cls}}$	1.00
	$\mathcal{L}_{\text{branch}}$	0.25
	$\mathcal{L}_{\text{supcon}}$	0.10
	$\mathcal{L}_{\text{edl}}$	0.05
	$\mathcal{L}_{\text{pre}}^{\text{real}}$	0.05
Forensic	$\mathcal{L}_{\text{cls}}$	1.00
	$\mathcal{L}_{\text{expert}}$	0.10
	$\mathcal{L}_{\text{supcon}}$	0.10
	$\mathcal{L}_{\text{edl}}$	0.05
	$\mathcal{L}_{\text{decorr}}$	0.05
Fusion	$\mathcal{L}_{\text{cls}}$	1.00
	$\mathcal{L}_{\text{supcon}}$	0.10
	$\mathcal{L}_{\text{stream}}$	0.05
	$\mathcal{L}_{\text{cons}}$	0.05

where $\mathcal{L}_{\text{cls}}$ supervises the final fused prediction, $\mathcal{L}_{\text{supcon}}$ regularizes the fused embedding, $\mathcal{L}_{\text{stream}}$ provides auxiliary supervision for stream-level predictions, and $\mathcal{L}_{\text{cons}}$ encourages consistency between individual stream probabilities and the final fused prediction.

Only the cross-stream Transformer, evidence projection, final classifier, embedding head, and auxiliary stream heads are optimized in this stage. All trainable Fusion components use a learning rate of 1×10^{-4}.

C.6 Loss Configuration

Table 3 summarizes the relative weights of the supervised objectives used in the three forgery-classification stages. The classification loss of each stage serves as the primary objective, while auxiliary terms are assigned smaller coefficients.

C.7 Optimization and Model Selection

Unless otherwise specified by the stage-specific optimization strategies above, all models are trained using AdamW with a weight decay of 1×10^{-4}. We use a batch size of 32 and train for at most 30 epochs with cosine learning-rate annealing. Automatic mixed-precision training is enabled, and the gradient norm is clipped to 5.0 for numerical stability.

Early stopping is applied with a patience of three epochs. An improvement is considered valid only when the monitored validation metric changes by at least 1×10^{-4}. For self-supervised Relation pre-training, the checkpoint with the lowest validation loss is selected. For all forgery-classification stages, model selection is based on the highest video-level ROC-AUC on the validation split.

The shared training settings are summarized in Table 4.

C.8 Video-level Evaluation

The test split is used only for final evaluation and does not participate in checkpoint selection.

A video generally produces multiple valid temporal windows. The model independently predicts the forgery probability of each window. Given N valid windows with probabilities $\{p_i\}_{i=1}^{N}$, the final video-level forgery probability is obtained by arithmetic averaging:

$$p_{\text{video}} = \frac{1}{N} \sum_{i=1}^{N} p_i. \tag{14}$$

Table 4: Shared training and optimization settings.

Hyperparameter	**Setting**
Window length	48 frames
Window stride	12 frames
Embedding dimension	128
Batch size	32
Optimizer	AdamW
Weight decay	1×10^{-4}
LR scheduler	Cosine annealing
Maximum epochs	30
Gradient clipping	5.0
Mixed precision	Enabled
Early-stopping patience	3 epochs
Minimum improvement	1×10^{-4}
Pre-training selection	Minimum validation loss
Classification selection	Maximum video-level ROC-AUC

We report video-level ROC-AUC, computed from the aggregated video probabilities and the corresponding ground-truth labels. This threshold-independent metric is used consistently across the main experiments and supplementary evaluations. For forgery-classification stages, checkpoints are selected using video-level ROC-AUC on the validation split, and the test split is reserved for final evaluation.

D Challenging Cross-Domain Analysis on AVLips

D.1 Evaluation Setting

We further investigate a challenging cross-domain setting on AVLips(Liu et al., 2024). We use a balanced subset consisting of the first 1,000 real and 1,000 fake videos and evaluate all methods under the same cross-dataset protocol used in our main experiments. In particular, MoCouple is trained on MMDF without access to AVLips during training.

Under this setting, MoCouple achieves 56.01% video-level AUC. Interestingly, LipDA(She et al., 2026), which similarly exploits relationships between lip motion and head pose, obtains a comparable AUC of 55.04%. These results contrast with the strong performance reported on AVLips when the training distribution is more closely associated with the target domain, suggesting that the degradation observed here should be interpreted primarily as a challenging cross-domain generalization problem rather than an inherent difficulty of detecting forgeries in AVLips.

D.2 Visual and Temporal Domain Shift

To better understand this behavior, we manually inspect videos from the evaluated AVLips subset. Compared with MMDF, many samples exhibit noticeably lower visual quality and substantial frame-to-frame spatial jitter. The latter is particularly visible in the background: the surrounding scene may shift across consecutive frames even when the facial subject exhibits no corresponding large viewpoint change. Since the released AVLips videos are already spatially cropped, the exact origin of this phenomenon cannot be determined from the available data alone and may be associated with characteristics of the provided preprocessing or cropping procedure.

Such temporal perturbations are particularly relevant to relation-based forgery detection. MoCouple models dependencies among vocal articulation, facial expression, and head motion, while LipDA similarly exploits the relationship between lip motion and head pose. Ideally, the extracted visual motion primarily represents intrinsic facial dynamics. Under frame-wise spatial instability, however, the observed motion can additionally contain non-semantic temporal variations:

$$\widetilde{\mathbf{M}}_t = \mathbf{M}_t^{\text{facial}} + \boldsymbol{\Delta}_t^{\text{jitter}}, \tag{15}$$

where $\boldsymbol{\Delta}_t^{\text{jitter}}$ represents temporal variation unrelated to genuine facial motion.

Because vocal articulation, expression, and head motion are extracted from different spatial regions and contextual scales, such perturbations need not affect their representations identically. They may therefore introduce relative deviations among the extracted motion features and shift their coupling distributions away from those learned on MMDF. We emphasize that this provides a plausible explanation consistent with our qualitative observations rather than establishing a causal relationship between spatial jitter and detection performance.

D.3 Stream-wise Analysis

The stream-wise results further characterize this domain shift. On AVLips, the Relation Stream achieves 58.16% AUC, whereas the Visual Forensic Stream obtains 49.21%. Their fusion results in 56.01% AUC.

The Relation Stream therefore retains more discriminative information than the Forensic Stream under this challenging setting, despite being affected by the cross-domain shift. Interestingly, LipDA obtains a similar overall performance of 55.04%. Since LipDA also models relationships between lip and head motion, the comparable degradation of two independently designed relation-based approaches suggests that motion-relation representations learned from a substantially different source distribution may be particularly sensitive to this type of domain shift.

Meanwhile, the near-chance performance of the Forensic Stream indicates an additional visual-domain mismatch. Consequently, combining the degraded forensic evidence with the Relation Stream reduces performance from 58.16% to 56.01%. This exposes a limitation of the current fusion mechanism: although it is designed to handle heterogeneous and conflicting evidence, severe out-of-domain degradation of an entire stream may not be fully suppressed.

Overall, the AVLips experiment highlights a distinction between generalization to *unseen generation mechanisms* and generalization across broader *visual and temporal domains*. MoCouple performs strongly against unseen talking-face generators in the primary experiments, whereas the MMDF-to-AVLips transfer additionally introduces substantial changes in visual quality, preprocessing characteristics, and temporal stability. The comparable degradation observed for LipDA further indicates that robustness of motion-relation representations across such domain shifts remains an important problem.

Future work may investigate explicit global-motion compensation, stabilization-aware motion representations, stronger stream reliability estimation, and training-time augmentation that simulates spatial jitter and other temporal preprocessing variations.