DIFFMOTION: DIFFERENTIAL FACIAL MOTION MODELING FOR ROBUST DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in deepfake generation have made forged facial videos increasingly realistic, challenging detectors that rely on appearance artifacts. We propose **DiffMotion**, a landmark-sequence framework that models facial motion as a complementary forgery cue. DiffMotion combines raw coordinates with bidirectional lag-1 and lag-2 differences, propagates them over a muscle-inspired sparse facial graph, and aggregates local temporal context using folded, overlapping windows. This design provides an interpretable account of which facial regions and motion relations contribute to a decision. Experiments on FaceForensics++ and Celeb-DF show competitive within-domain performance and strong resilience to controlled compression and landmark perturbations; they also reveal a substantial cross-dataset gap. Notably, DiffMotion maintains an accuracy of over 86.75% under severe noise and high frame-drop rates, significantly outperforming existing methods.We therefore position DiffMotion as an anatomically motivated and degradation-resilient detector, while treating domain generalization as an important limitation rather than a solved problem.

1 INTRODUCTION

Advances in deep generative models make synthetic facial content Thies et al. (2020); Deepfakes (2018); Li et al. (2020a); GitHub, Inc. (2025) increasingly realistic, and forged facial videos—deepfakes—pose severe security and ethical risks, making reliable detection critical Rossler et al. (2019); Agarwal et al. (2020).

Existing detectors analyze texture, temporal consistency, or pretrained features, but often overfit dataset-specific artifacts Agarwal et al. (2020) and pixel-level appearance or coarse geometry Yang et al. (2019), overlooking fine-grained dynamics and structured facial motion.

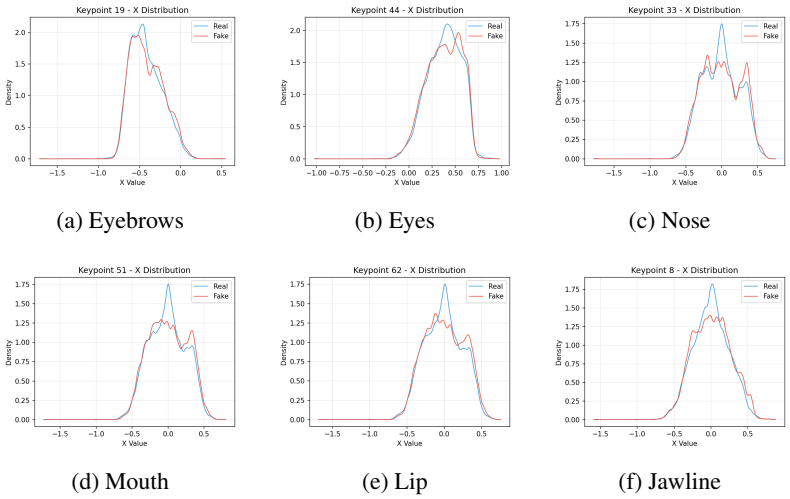
We test whether small deviations persist in landmark trajectories and inter-region coordination even when frames look plausible, since real facial motion is structured by muscles Ekman & Friesen (1978) and temporal behavior complements appearance Agarwal et al. (2019); we combine (i) multi-lag bidirectional differences, (ii) a sparse muscle-inspired facial graph, and (iii) folded temporal windows with temporal convolution.

On FaceForensics++ and Celeb-DF Rossler et al. (2019); Li et al. (2020b), DiffMotion achieves competitive within-domain and cross-manipulation performance and resilience to compression and landmark perturbations, while cross-dataset transfer remains limited; degradation robustness and domain generalization are treated as distinct throughout.

To this end, we propose DiffMotion, a landmark-centric detector, with three contributions:

- **Keypoint differential features:** raw coordinates plus bidirectional lag-1/lag-2 differences expose motion variations invisible in static geometry.
- **Muscle-group topology:** a sparse, muscle-inspired graph over 68 landmarks constrains message passing to plausible co-motion paths.
- **Folded temporal windows:** overlapping three-frame windows fold local context into spatial attention, followed by temporal convolution.

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.



(a) Eyebrows (b) Eyes (c) Nose

(d) Mouth (e) Lip (f) Jawline

Figure 1: Representative KDE-based landmark distribution comparison across facial regions. Real faces exhibit smooth unimodal patterns, while synthetic faces show multi-modal or shifted densities, especially in expressive regions.

2 Related Work

2.1 Deepfake Generation and Detection

Face manipulation has progressed from visual-only identity swapping Nirkin et al. (2019); Li et al. (2020a); GitHub, Inc. (2025); Deepfakes (2018) and expression reenactment Thies et al. (2020) to audio-visual synthesis that drives lips and talking faces from speech Prajwal et al. (2020); Song et al. (2022); Chen et al. (2019); Vougioukas et al. (2020). Detection correspondingly targets spatial artifacts—CNN baselines Rossler et al. (2019), frequency-aware cues Kashiani et al. (2025), patch source-consistency Zhao et al. (2021b), and one-class anomaly scores Khalid & Woo (2020), often on transformer backbones Liu et al. (2021)—or multimodal speech–motion–affect inconsistencies Astrid et al. (2025); Mittal et al. (2020). Both directions rely on appearance or cross-modal artifacts rather than facial dynamics itself, which provides complementary evidence.

2.2 Landmark-Based Facial Dynamics

Landmarks compactly represent facial dynamics and have supported behavioral cues such as eye, mouth, and head movement Agarwal et al. (2020), head-pose consistency Yang et al. (2019), and region-specific lip motion Haliassos et al. (2021); expression analysis further organizes facial motion into regions and action units Li et al. (2018). Existing landmark detectors, however, mostly use coordinates through distances, statistics, or per-region features, leaving the coordinated motion between landmarks largely unmodeled. Since real facial motion exhibits strong inter-landmark correlation, DiffMotion adopts a sparse graph of selected co-motion relationships as its structural prior.

2.3 Spatio-Temporal Modeling

Manipulation traces appear at multiple temporal scales, from frame-to-frame deviations to longer-range dynamics such as expression consistency. Hierarchical or windowed video models Fan et al. (2021); Liu et al. (2021) balance local and global context better than full attention, but structure-aware multi-scale modeling of landmark dynamics remains unexplored; DiffMotion combines graph-constrained spatial attention with folded local windows and temporal convolution.

3 MOTIVATION

As generators improve, we ask whether landmark dynamics retain measurable real–fake differences even when frames look plausible, treating motion as a complementary cue without assuming natural motion cannot be synthesized. Real landmarks follow smooth, unimodal distributions reflecting anatomical variability, while synthetic landmarks show shifted peaks, suppressed modes, or spurious multi-modality, particularly in muscle-driven regions—motivating *muscle-informed representation learning*.

3.1 PROBABILITY DISTRIBUTION ANALYSIS OF FACIAL LANDMARKS

Per-region 1D Gaussian KDEs over landmarks grouped into five anatomical regions (eyebrows, eyes, nose, mouth, jawline) show smooth, unimodal real distributions versus shifted or multi-modal synthetic patterns (Figure 1), strongest in muscle-driven lower-face regions.

3.2 STATISTICAL CONSISTENCY ANALYSIS OF TEMPORAL DYNAMICS

We compare real and synthetic trajectories and their first-order differentials using Dynamic Time Warping (DTW) and lag-1 autocorrelation, with $\mathbf{x} = \{x_t\}_{t=1}^{T}$ a trajectory and $\Delta\mathbf{x}$ its first-order differential.

Dynamic Time Warping (DTW). We align each sequence to an average real-motion template using DTW:

$$\mathrm{DTW}(\mathbf{x}, \mathbf{y}) = \min_{\pi} \sum_{(i,j)\in\pi} \|x_i - y_j\|_2, \tag{1}$$

where π is a monotonic alignment path. Raw trajectories already differ significantly between real and synthetic sequences (Real: 209.3, Fake: 219.8; $t = -3.117$, $p = 0.00188$; KS: $p = 0.00277$), and differential trajectories increase separability further (Real: 17.3, Fake: 16.6; $t = 3.08$, $p = 0.002$; KS: $p = 0.043$).

Autocorrelation of Temporal Persistence. To capture short-term temporal dependency, we compute the lag-1 autocorrelation for each landmark coordinate:

$$\rho(\mathbf{x}) = \frac{\sum_{t=1}^{T-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}. \tag{2}$$

Raw trajectories show little difference between real and synthetic videos (Real: 0.926, Fake: 0.925; $t = 0.3637$, $p = 0.7161$; KS: $p = 0.0165$), but the gap grows substantially after differencing (Real: 0.041, Fake: 0.021; $t = 4.003$, $p = 6.8\times10^{-5}$; KS: $p = 1.08\times10^{-4}$), revealing weaker temporal continuity in synthetic motion and motivating *differential motion encoding*.

3.3 INTER-LANDMARK CORRELATION PATTERNS

Pearson correlations among all 68 landmarks capture relational coordination: real motion shows organized block structures consistent with coordinated muscle activation (eyes, eyebrows, mouth), while synthetic motion shows weaker, diffuse correlations (Figure 2).

From Correlation Patterns to Facial Anchor Graph. Because landmarks in the same anatomical region move together and neighboring structures provide complementary constraints, we encode these dependencies in a sparse facial anchor graph.

Let the facial landmark graph be defined as

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{3}$$

with $\mathcal{V} = \{v_0, \ldots, v_{67}\}$ the 68 landmarks and $\mathcal{E}$ connecting anatomically related landmarks; the topology is derived from facial anatomy and observed correlations rather than full connectivity.



Figure 2: Pearson correlation heatmaps for real and synthetic facial motion sequences. Real videos exhibit structured block patterns reflecting coordinated muscle dynamics, whereas synthetic videos show weakened and diffuse inter-landmark correlations. The difference map highlights structural inconsistencies in generated temporal dynamics.

The topology preserves natural contours and local structure: the jawline chains landmarks 0–16, the eyebrows chain 17–26, the nasal bridge connects 27–30, each eye forms a closed structure (36–41 and 42–47), and the mouth uses nested outer (48–59) and inner (60–67) lip contours.

A small number of cross-region edges link eye landmarks to mouth-side anchors (36, 40, 41 to 48; 45, 46, 47 to 54) and join the lip contours at the corners (48 to 60, 54 to 64), capturing coordinated deformation without full connectivity.

Formally, the edge set can be decomposed into several anatomical components:

$$\mathcal{E} = \mathcal{E}_{\text{jaw}} \cup \mathcal{E}_{\text{brow}} \cup \mathcal{E}_{\text{nose}} \cup \mathcal{E}_{\text{eye}} \cup \mathcal{E}_{\text{mouth}} \cup \mathcal{E}_{\text{cross}}, \tag{4}$$

Each component corresponds to one facial structure: adjacent landmarks are connected along sequential contours, closed structures such as the eyes and lips form cyclic graphs, and the resulting sparse graph is illustrated in Figure 3.

Together, synthetic faces show distributional shifts in muscle-driven regions, differential dynamics amplify real–fake separability, and inter-landmark correlations reveal disrupted coordination, motivating explicit modeling of temporal differentials and muscle-informed spatial relationships.

4 Method

DiffMotion combines a multi-lag differential encoder, a muscle-inspired facial graph, and folded-window spatial attention with temporal convolution (Figure 4), operating on 68 two-dimensional landmarks from one face per video.

4.1 Landmark Representation and Differential Encoding

Let $\mathbf{X} \in \mathbb{R}^{C\times T\times V}$ denote a sequence of $V = 68$ landmarks over T frames with $C = 3$ channels: normalized x and y, plus a constant auxiliary channel with zero temporal difference. For lag $k \in \{1, 2\}$ we compute

$$\Delta^{(k,+)}\mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-k}, \qquad \Delta^{(k,-)}\mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t+k}. \tag{5}$$

These bidirectional first differences (not derivatives) are zero-padded at unavailable indices—k frames prepended/appended—without circular padding or boundary masks.

The raw sequence is projected by a 1×1 convolution from C to 32 channels, and each of the four difference branches is projected from C to 8 channels. Their concatenation

$$\mathbf{M} = \left[\phi_0(\mathbf{X}), \phi_{1,+}(\Delta^{(1,+)}\mathbf{X}), \phi_{1,-}(\Delta^{(1,-)}\mathbf{X}), \phi_{2,+}(\Delta^{(2,+)}\mathbf{X}), \phi_{2,-}(\Delta^{(2,-)}\mathbf{X})\right] \in \mathbb{R}^{64\times T\times V} \tag{6}$$

preserves static geometry while exposing short-range directional changes.

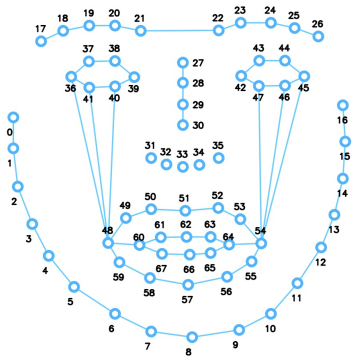


Figure 3: The sparse facial graph used by DiffMotion. Nodes follow the standard 68-landmark indexing scheme. Edges encode facial contours and selected cross-region co-motion relations; the graph is muscle-inspired rather than a complete anatomical model.

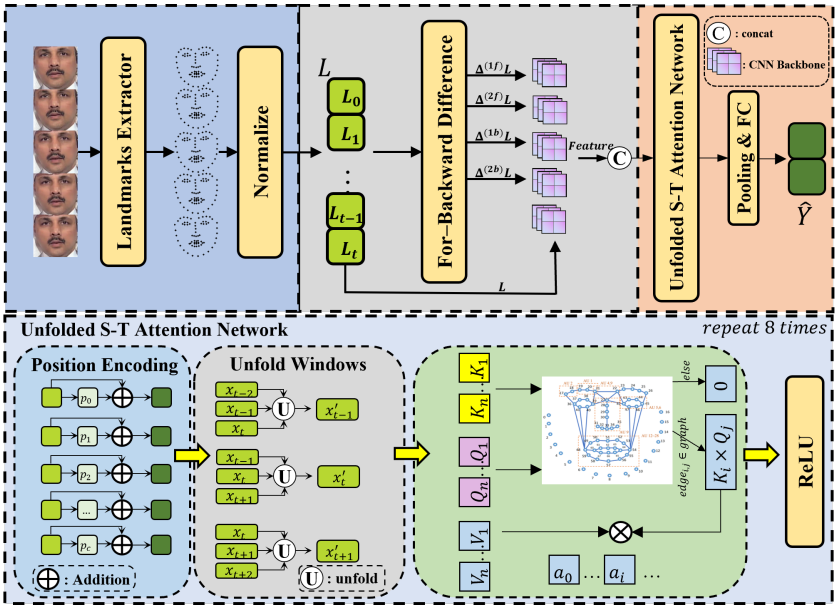


Figure 4: Overview of DiffMotion. Raw landmarks and four bidirectional lagged differences are embedded and concatenated. Graph-masked spatial attention aggregates each three-frame folded window, temporal convolutions model the resulting sequence, and global pooling followed by a linear classifier predicts real or fake.

4.2 Muscle-Inspired Facial Graph

We define $G = (\mathcal{V}, \mathcal{E})$ over the standard 68-point layout: jaw 1–17, eyebrows 18–27, nose 28–36, right eye 37–42, left eye 43–48, outer lip 49–60, and inner lip 61–68. Local edges follow the corresponding contours. We additionally use the cross-region edges

$$\begin{aligned}\mathcal{E}_{\text{cross}} = \{&(37,49),(42,49),(41,50),(48,54),(47,55),(46,55),\\ &(32,49),(33,51),(34,52),(35,53),(36,55),(63,67),(58,9),(55,65)\}.\end{aligned} \tag{7}$$

This sparse edge set is an engineering approximation of facial co-motion rather than an exhaustive anatomical model.

Let A be the binary undirected adjacency and I the identity matrix. We use symmetric normalization

$$\widetilde{A} = D^{-1/2}(A + I)D^{-1/2}. \tag{8}$$

Graph-distance subsets built from powers of the binary adjacency mask spatial attention; the reported FF++ configuration uses one- and two-hop neighborhoods with self-connections, so a landmark attends only to permitted graph neighborhoods. A learned global term and a data-dependent sigmoid gate keep the topology constraining without fixing all interaction weights.

4.3 Folded-Window Spatial Attention and Temporal Modeling

For every time index, we extract a length-$L = 3$ temporal window with stride one and one zero frame of padding at each temporal boundary. The unfold operator concatenates the three frames along the landmark dimension,

$$\mathcal{U}(\mathbf{Z})_t = [\mathbf{Z}_{t-1}; \mathbf{Z}_t; \mathbf{Z}_{t+1}] \in \mathbb{R}^{C' \times 3V}, \tag{9}$$

so the number of temporal positions remains T while the spatial support changes from V to $3V$.

For each graph subset, values are the unfolded features; keys are projected from current-frame V features and queries from the $3V$ folded features. The affinity is scaled by the channel dimension times T, masked by the repeated graph subset, softmax-normalized over the V source landmarks, modulated by a learned difference gate, and aggregated back to V landmarks; the subsets act as parallel branches without multi-head QKV splitting.

Eight blocks each apply the graph-masked spatial module followed by a temporal convolution (kernel size 7, padding 3). The block channel/stride schedule is

$$(64, 64, 1)^2, (64, 128, 2), (128, 128, 1), (128, 256, 2), (256, 256, 1)^3, \tag{10}$$

where each tuple gives input channels, output channels, and temporal stride. BatchNorm, LeakyReLU with slope 0.1, feed-forward convolutions, and residual paths are used with zero attention/feature dropout; temporal aggregation comes from the folded local context and temporal convolutions, not temporal self-attention or positional encoding.

4.4 Classification and Complexity

After the eight blocks, global average pooling over time and landmarks produces a 256-dimensional representation followed by a linear $256 \to 2$ classifier trained with cross-entropy loss. The reported two-subset model contains 4,121,922 trainable parameters.

5 Experiments

5.1 Datasets

We evaluate on two widely used deepfake datasets:

- **FaceForensics++ (FF++) C23/C40**: C23 for main training and within-quality evaluation; C40 for transfer to stronger compression; the DF, F2F, FS, and NT subsets for cross-manipulation evaluation.
- **Celeb-DF (V1/V2)**: a high-quality benchmark collected from online videos, used to test robustness against realistic synthesis.

5.2 Evaluation Protocol

We adopt a **train-once, test-everywhere** paradigm: DiffMotion is trained on FF++ (C23) with fixed preprocessing and evaluation settings, then tested across domains, manipulations, and quality levels, which are reported separately; baseline entries from prior work are identified as reported results rather than unified reimplementations.

Evaluation Metric. AUC is the primary metric; one deterministic 128-frame clip per video is sampled, pooled over time and landmarks, and evaluated without frame-level voting or test-time augmentation.

Table 1: Cross-manipulation generalization on FF++. Models are trained on a single manipulation subset (DF/F2F/FS/NT) and evaluated on all four subsets. AUC (%) is reported.

Training Set	Method	DF	F2F	FS	NT	Avg.
DF	DCLSun et al. (2022)	**99.98**	77.13	61.01	75.01	78.28
	UIA-ViTZhuang et al. (2022)	99.29	74.44	53.89	70.92	74.64
	WATCHERWang et al. (2024)	99.68	75.50	79.51	**83.66**	84.59
	FreqDebiasKashiani et al. (2025)	99.82	**86.45**	75.92	**88.45**	**87.66**
	DiffMotion (Ours)	99.62	70.18	**80.57**	71.82	80.55
F2F	DCLSun et al. (2022)	**91.91**	99.21	59.58	66.67	79.34
	UIA-ViTZhuang et al. (2022)	83.39	98.32	68.37	67.17	79.31
	WATCHERWang et al. (2024)	85.73	**99.82**	69.38	74.16	82.27
	FreqDebiasKashiani et al. (2025)	99.41	99.44	74.37	76.46	87.42
	DiffMotion (Ours)	85.24	99.17	**85.66**	**82.38**	**88.11**
FS	DCLSun et al. (2022)	74.80	69.75	99.90	52.60	74.26
	UIA-ViTZhuang et al. (2022)	81.02	66.30	99.04	49.26	73.91
	WATCHERWang et al. (2024)	84.52	75.06	**99.95**	58.81	79.59
	FreqDebiasKashiani et al. (2025)	83.76	**78.93**	99.78	63.48	81.49
	DiffMotion (Ours)	**89.36**	58.72	99.40	**89.36**	**84.21**
NT	DCLSun et al. (2022)	91.23	52.13	79.31	98.97	80.41
	UIA-ViTZhuang et al. (2022)	79.37	67.89	45.94	94.59	71.95
	WATCHERWang et al. (2024)	89.67	72.85	69.96	98.58	82.77
	FreqDebiasKashiani et al. (2025)	**92.35**	**74.61**	83.24	99.83	87.50
	DiffMotion (Ours)	90.36	62.80	**99.75**	**99.84**	**88.18**

5.3 Implementation Details

Landmark extraction and normalization. Per frame we extract 68 2D landmarks with OpenFace FeatureExtraction (2D only). The bounding box is enlarged by $1.2\times$, stabilized by the mean box size over the video (centers below 1.42 pixels retain the previous center), and resized to 400×400; coordinates are centered at $(200, 200)$, divided by 200, and frames without a valid face box are skipped.

Forward Lucas–Kanade tracking (15×15 window, three pyramid levels) with a backward consistency check and per-landmark Kalman fusion (process variance 0.3) reduces jitter; a point is not fused when optical flow fails or its feedback distance exceeds 2 pixels. One face per video is assumed.

Sequence construction and sampling. Each frame (136 values) is reshaped to two coordinate channels over 68 landmarks plus a constant third channel of 0.1 for skeleton compatibility. Sequences are truncated at 1,400 frames and zero-padded to 1,500; training samples 150 frames with replacement, sorts them, and selects a random contiguous 128-frame segment, while evaluation uses 150 uniform samples to take the centered 128-frame segment.

Optimization. We train with SGD (momentum 0.9, Nesterov), weight decay 5×10^{-4}, batch size 32, learning rate 0.1 with a five-epoch linear warm-up and $0.1\times$ drops at epochs 60/90, for up to 1,200 epochs; no label smoothing or mixup, Python/NumPy/PyTorch seeds of 1, deterministic cuDNN, and two GPUs.

5.4 Comprehensive Evaluation

Unless noted, DiffMotion is trained on **FaceForensics++ (C23)** and evaluated across domains and manipulations; baseline entries from original reports make the tables contextual comparisons rather than a same-codebase benchmark.

Cross-manipulation Performance. Each model is trained on one FF++ manipulation (DF/F2F/FS/NT) and tested on all four (Table 1). Most detectors are near-perfect on the seen ma-

nipulation but degrade on unseen ones (e.g., DCLSun et al. (2022): 99.98 on DF to 61.01 on FS), indicating reliance on manipulation-specific artifacts.

DiffMotion transfers more stably: F2F-trained reaches 85.66/82.38 on unseen FS/NT (best average 88.11); FS- and NT-trained give the best averages of 84.21 and 88.18 with strong FS-to-NT transfer (89.36); and it achieves the best FS AUC (80.57) when DF-trained despite weak DF-to-FS transfer overall, suggesting behavior-aware representations capture manipulation-invariant cues.

Cross-dataset Performance. Models trained on FF++ (C23) are evaluated on Celeb-DF V1/V2 (Table 2). All methods degrade under this domain shift (compression, generation pipelines, video statistics); DiffMotion reaches 97.9 within-domain but 66.70/59.62 on Celeb-DF, showing motion statistics vary across datasets and motivating the robustness analysis below.

Table 2: Comprehensive deepfake detection results (AUC %). All models are trained on FF++ (C23). The first column reports within-domain performance on FF++, while the last two columns evaluate cross-dataset generalization.

Method	**C23**	**Celeb-DF V1**	**Celeb-DF V2**
XceptionRossler et al. (2019)	96.8	82.7	72.5
Multi-attZhao et al. (2021a)	96.4	67.4	/
LipForensicsHaliassos et al. (2021)	**98.3**	**85.1**	**76.2**
FreqDebiasKashiani et al. (2025)	97.5	**87.5**	**83.6**
DiffMotion (Ours)	97.9	66.70	59.62

5.5 Degradation and Compression Robustness Evaluation

Temporal and spatial degradation robustness. To emulate missing temporal observations and landmark localization error, corruptions are applied to test landmarks: a proportion of frames is retained and linearly interpolated back to the original length (90%, 70%, 50%, 30% denote retained frames), and Gaussian noise is added to normalized coordinates for spatial degradation. The ratio directly controls observed motion information in our representation, whereas in physiological methods such as UMCL it mainly changes rPPG temporal resolution; results are in Table 3.

Table 3: Robustness evaluation under different signal degradations on FF++. Sampling ratio denotes retained frames followed by linear interpolation, and perturbation injects Gaussian coordinate noise. Degradation is applied at test time. AUC (%) is reported.

	UMCL Lai et al. (2026)		**DiffMotion (Ours)**	
Degradation	FF++ (c23)	FF++ (c40)	FF++ (c23)	FF++ (c40)
No degradation	96.29	79.72	**99.65**	**94.77**
Sampling ratio 90%	96.54	78.74	**99.57**	**93.03**
Sampling ratio 70%	95.84	81.22	**99.37**	**92.69**
Sampling ratio 50%	92.81	80.04	**98.93**	**91.49**
Sampling ratio 30%	85.69	75.56	**97.24**	**86.75**
Perturbation $\sigma = 0.01$	93.77	74.72	**98.37**	**92.41**
Perturbation $\sigma = 0.05$	75.68	61.96	**95.98**	**89.86**

Evaluation protocol. Perturbations are applied only to test landmarks of medium- and heavily compressed FF++ videos; since UMCL uses a different physiological representation, the comparison is cross-method context rather than a strictly modality-equivalent intervention.

Discussion. DiffMotion stays robust under temporal degradation (at 30% sampling, C23 AUC drops from 99.65 to 97.24) and outperforms UMCL under the reported settings; spatial perturbations degrade more as localization noise propagates to alignment and motion estimation, but accuracy remains competitive (95.98 at $\sigma = 0.05$).

5.6 Ablation Study

To assess each design choice, we ablate on FF++ (C23):

- **w/o Bidirectional Differencing**: only forward differencing is retained.
- **w/o Temporal Differencing**: temporal differencing is removed, leaving static keypoint positions.
- **Full Graph**: the facial topology is replaced by a fully connected graph.

Results (Table 4) show temporal differencing matters most: removing it drops AUC from 99.65 to 70.58 (-29.07), as static geometry cannot capture manipulation's temporal inconsistencies.

Table 4: Ablation study on FF++ (C23). Removing temporal differencing leads to a large performance drop, while bidirectional motion modeling and facial topology both contribute to the final gain.

Variant	**FF++ C23**
w/o Bidirectional Differencing	96.35
w/o Temporal Differencing	70.58
Full Graph (no facial topology)	94.15
DiffMotion (full model)	**99.65**

Removing bidirectional differencing yields 96.35 (-3.30): forward differencing captures most of the signal, but both directions provide complementary dynamics. A fully connected graph yields 94.15 (-5.50), as unrestricted propagation adds irrelevant interactions between distant landmarks. Together, temporal differencing and the topology-aware graph are essential and achieve the best performance.

6 LIMITATION

DiffMotion is resilient under the reported degradations, but cross-dataset generalization remains limited: Celeb-DF transfer shows that landmark distributions and motion statistics encode dataset-specific factors (capture conditions, compression, alignment); degradation robustness and domain generalization are therefore distinct claims.

The method depends on reliable landmark localization: frames without a valid OpenFace face are skipped and one face per video is assumed, so pose, occlusion, blur, or compression can remove frames or introduce trajectory errors; our perturbations cover only part of this failure space. The sparse topology is an engineering prior rather than a complete biomechanical model, and the folded-window module captures only three-frame local context with temporal convolution.

Future work should compare landmark detectors, quantify failures, and evaluate domain normalization or multimodal fusion with same-pipeline corruptions, multiple seeds, and explicit success/failure cases.

7 CONCLUSION

We presented DiffMotion, a landmark-based detector that combines raw facial geometry with bidirectional lagged differences, a sparse muscle-inspired graph, and folded local temporal windows. The results support temporal motion and structured landmark relations for within-domain detection, selected cross-manipulation transfer, and controlled degradation robustness, while the weaker Celeb-DF transfer shows that motion representations remain sensitive to domain shift. DiffMotion is therefore an interpretable, degradation-resilient behavioral detector rather than a complete solution to cross-dataset generalization; improving detector reliability and domain-invariant motion representations are central future directions.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for editing the research paper to improve readability and clarity, suggesting a structure for the paper, and assisting with the creation or editing of software code.

We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in writing proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing the core research methods, translating research materials, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting experimental results. The remaining required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for brainstorming, suggesting experimental parameters, improving the readability of the manuscript, identifying relevant literature, summarizing existing literature, and formatting references. All AI-assisted work was reviewed and verified by the authors. AI-assisted code was manually inspected, tested, and verified for correctness by the authors. AI-assisted text was carefully reviewed against the original research content, experimental results, and cited literature to ensure that no unsupported claims or factual errors were introduced. The authors take full responsibility for the final content of this work, including all text, claims, methods, results, and artifacts produced with the aid of generative AI.

REFERENCES

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, New York, NY, USA, 2020. IEEE. doi: 10.1109/WIFS49906.2020.9360904.

Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. Audio-visual deepfake detection with local temporal inconsistencies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889087.

Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7824–7833, Long Beach, CA, USA, 2019. IEEE. doi: 10.1109/CVPR.2019.00802.

Deepfakes. faceswap github, 2018. URL `https://github.com/deepfakes/faceswap`.

P Ekman and W V Friesen. Facial action coding system (facs): a technique for the measurement of facial actions. *Rivista Di Psichiatria*, 47(2):126–38, 1978.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6804–6815, Montreal, Canada, 2021. IEEE. doi: 10.1109/ICCV48922.2021.00675.

GitHub, Inc. Faceswap topic. `https://github.com/topics/faceswap`, 2025. Accessed: January 4, 2025.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection, 2021. URL `https://arxiv.org/abs/2012.07657`.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8775–8785, June 2025.

Hasam Khalid and Simon S. Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2794–2803, Seattle, WA, USA, 2020. IEEE. doi: 10.1109/CVPRW50498.2020.00336.

Ching-Yi Lai, Chih-Yu Jian, Pei-Cheng Chuang, Chia-Ming Lee, Chih-Chung Hsu, Chiou-Ting Hsu, and Chia-Wen Lin. Umcl: Unimodal-generated multimodal contrastive learning for cross-compression-rate deepfake detection. *International Journal of Computer Vision*, 134(1):40, 2026.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2020a. URL https://arxiv.org/abs/1912.13457.

Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2583–2596, 2018. doi: 10.1109/TPAMI.2018.2791608.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: a large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213, 2020b. doi: 10.1109/CVPR42600.2020.00327.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, Montreal, Canada, 2021. IEEE.

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*, MM '20, pp. 2823–2832, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413570. URL https://doi.org/10.1145/3394171.3413570.

Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment, 2019. URL https://arxiv.org/abs/1908.05932.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. 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. ACM, 2020. doi: 10.1145/3394171.3413532.

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 (ICCV)*, pp. 1–11, Seoul, Korea, 2019. IEEE.

L Song, W Wu, C Qian, Y Liu, and C. C Loy. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598, 2022.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2316–2324, Jun. 2022. doi: 10.1609/aaai.v36i2.20130. URL https://ojs.aaai.org/index.php/AAAI/article/view/20130.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos, 2020. URL https://arxiv.org/abs/2007.14808.

K Vougioukas, S Petridis, and M Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020.

Yuan Wang, Chen Chen, Ning Zhang, and Xiyuan Hu. Watcher: Wavelet-guided texture-content hierarchical relation learning for deepfake detection. *Int. J. Comput. Vision*, 132(10):4746–4767, May 2024. ISSN 0920-5691. doi: 10.1007/s11263-024-02116-5. URL https://doi.org/10.1007/s11263-024-02116-5.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, Brighton, UK, 2019. IEEE. doi: 10.1109/ICASSP.2019.8683164.

Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Weiming Zhang, Dongdong Chen, and Nenghai Yu. Multi-attentional deepfake detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2185–2194, 2021a. doi: 10.1109/CVPR46437.2021.00222.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15003–15013, Montreal, Canada, 2021b. IEEE. doi: 10.1109/ICCV48922.2021.01475.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 391–407. Springer, 2022.