SYNDROMEFLOW: JOINT TEMPORAL-WARP AND RECOMPOSITION INFERENCE FOR AUDIOVISUAL WATERMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Replacing or replaying watermarked segments can alter a video while leaving the embedded watermarks intact. Benign audio delays and clock drift make these edits harder to localize: tampering can bias alignment, and alignment errors can in turn create false tamper evidence. We propose SyndromeFlow, an audiovisual watermarking framework that jointly infers alignment and sparse tampering under a shared Bayesian posterior. Our code links segments through parity checks, so copying a segment with its watermark can break consistency with surrounding content. The receiver records failed checks as a syndrome and scores how candidate alignments and edited regions jointly explain it. Local tamper evidence thus refines alignment, while posterior averaging produces separate audio and video localization maps. Compared with a matched receiver that fixes alignment using the pilot and path prior, SyndromeFlow increases audio AUPRC from 0.586 to 0.637, reduces mean absolute alignment error from 2.2 ms to 0.9 ms, and maintains video AUPRC above 0.99. A controlled comparison of the inference steps attributes most of the improvement to tamper evidence revising alignment; averaging contributes further localization gains when alignment remains uncertain. In a complete-system comparison on the same speaking-person benchmark, each method embeds its own watermark before local copying. SyndromeFlow achieves video AUPRC of 0.994, compared with 0.086 for the strongest of five image-watermark baselines.

1 INTRODUCTION

During publication and subsequent verification, internet videos often undergo multiple rounds of resampling. Such processing may introduce slight temporal offsets between audio and video and cause their relative timing to drift gradually over time. Local content may also undergo replacement or duplication. Proactive authentication methods embed evidence at publication to support subsequent detection and localization of these edits. This evidence relates received regions to the published content and supports integrity verification as editing methods and processing pipelines change.

Prior work has explored active authentication from several perspectives. Early studies focused on fragile and semi-fragile watermarking (Chen & Wornell, 2001; Rigoni et al., 2016; Alattar et al., 2020). Recent methods have extended this line to learned image, video, and audiovisual authentication (Zhang et al., 2024a; Sander et al., 2025; Hu et al., 2025; Zhang et al., 2024b; Kim et al., 2025; Zhang et al., 2025; Si et al., 2026). These methods progressively improve local tamper detection and cross-modal verification. Related studies further incorporate temporal information: Alattar et al. (Alattar et al., 2020) encode frame-level indices to detect insertion, deletion, and reordering, while LAVA (Zeng et al., 2026) estimates audiovisual offsets before score fusion. Natural audiovisual features have also been used to estimate temporal correspondence (Chung & Zisserman, 2016; Wang et al., 2020; Fernandez-Labrador et al., 2024). Together, these studies establish local authentication and temporal alignment as key components of audiovisual verification.

We study local replacement from another marked source and replay from an earlier interval of the same marked video. These local recompositions copy the embedded watermark together with the content. When benign temporal deformation and such editing coexist, similar authentication anomalies can have different causes. Unedited audio delayed by processing may fail checks tied to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

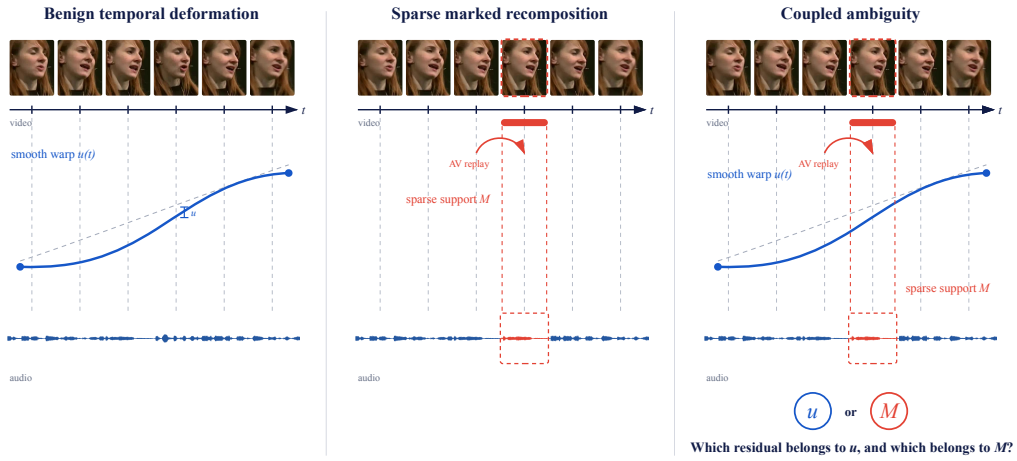

Figure 1: Alignment–localization ambiguity. Blue denotes a benign temporal deformation u; red denotes a one-second marked AV replay with sparse support M. Their competing explanations motivate joint inference.

the original timeline. Conversely, a replaced or duplicated segment can alter alignment evidence, causing the receiver to interpret editing as a timing change.

Temporal alignment and tamper localization therefore depend on each other. The alignment path determines which audio segments are checked with each video region. Judging a region as edited changes which checks should influence that path. A receiver that fixes alignment under an authentic-content assumption can be biased by local edits, especially when timing evidence is weak. Verification along the resulting path can spread a short edit into broader inconsistencies. Figure 1 illustrates this ambiguity and the central question of this paper: how can shared authentication evidence jointly determine temporal alignment and edited regions?

This observation motivates **SyndromeFlow**. At publication time, a keyed parity graph connects content states from separated video regions, audio bins, and times. At verification time, a temporal path reindexes the received audio states, and a sparse support variable identifies where editing may have occurred. The embedded parity measurements provide content-binding observations for inferring these changes. We score their combinations under one probability model and retain posterior marginals for alignment and localization. Our physical receiver operates on watermarks embedded in real AV media and extracted after editing and codec processing. It uses a time-varying pilot to propose smooth monotone paths and local visual evidence to propose supports. A controlled graph study examines belief propagation for larger path-conditional parity graphs.

Our contributions are:

- We formulate AV watermark authentication with temporal alignment and sparse tampering coupled as latent variables under benign timing changes.
- We design an AV parity representation that provides distributed content-binding evidence for joint inference.
- We develop a syndrome-based receiver that scores path–support hypotheses and marginalizes their posterior for audio and video localization.
- Complete-system comparisons and matched physical experiments quantify the gains from path revision and marginalization, and relate their contributions to alignment uncertainty.

2 Related Work

2.1 Active authentication and coded localization

QIM provides a principled family for embedding digital information in a host signal (Chen & Wornell, 2001). Rigoni et al. (Rigoni et al., 2016) apply modified QIM to AV content and localize spatial, global, and temporal tampering. Alattar et al. (Alattar et al., 2020) combine local hashes with changing segment and frame indices to expose insertion, deletion, reordering, track mismatch, and watermark copying. Other semi-fragile schemes use randomized nonlocal dependencies to make a transplanted region accountable to distant host content (Li & Yuan, 2006; Liu & Steinebach, 2006). Syndrome decoders and error-locating codes further connect authentication to structured inference (Lin et al., 2007; Lefèvre et al., 2022). This line supplies content binding and localization on a publishing timeline; benign changes of that timeline remain external to the decoding model.

2.2 Learned localized and audiovisual watermarks

Learned watermarking supports localized verification. EditGuard and OmniGuard combine copyright protection with tamper localization (Zhang et al., 2024a; 2025b). WAM decodes marked regions (Sander et al., 2025), while MaskWM learns masked embedding (Hu et al., 2025). ReImage embeds shuffled content for image recovery (Kim & Seo, 2026); AudioSeal localizes audio watermarks (San Roman et al., 2024). DiM supports variable payloads and temporal-order recovery (Meng et al., 2026).

AV systems distribute authentication evidence across modalities. V2A-Mark provides visual and audio localization together with copyright recovery (Zhang et al., 2024b). CMW stores information for authentic audio recovery inside visual frames (Kim et al., 2025). SepVAMark supports source tracing and deepfake detection through audiovisual fusion (Zhang et al., 2025), while mAVE binds the latent state of joint audio–video generation (Si et al., 2026). LAVA adds tempo correction and searches offsets before confidence-weighted score fusion (Zeng et al., 2026). Across these designs, alignment supplies an input to authentication; local support evidence enters after a path or offset has been selected.

2.3 Temporal correspondence and robust matching

SyncNet estimates speech AV lag from natural features (Chung & Zisserman, 2016). AlignNet learns dense correspondence for nonuniform timing (Wang et al., 2020), and DiVAS handles dynamic audio and video rates (Fernandez-Labrador et al., 2024).

Soft-DTW marginalizes smoothly over alignment costs (Cuturi & Blondel, 2017), and Drop-DTW aligns shared content while assigning unmatched observations to drop states (Dvornik et al., 2021). These algorithms couple a path to local matching states. SyndromeFlow evaluates paths and supports on a shared, host-bound parity graph whose factors span times, spatial regions, and modalities. Within the joint posterior, support hypotheses inform the path distribution, addressing the alignment–localization ambiguity studied here.

3 Method

Figure 2 summarizes the publisher and receiver. A shared parity graph carries the correspondence evidence into joint inference, allowing temporal alignment and local recomposition to inform each other.

3.1 Authentication under admissible temporal deformation

Let $X = (X_V, X_A)$ be a published AV work and $Y = (Y_V, Y_A)$ the received media, from which all decoder observations are computed. We treat smooth monotone temporal deformations as benign within a calibrated family covering offset, affine drift, and low-frequency warp. The physical receiver represents a candidate deformation by the path $\phi_u(t) = t + u(t)$ from a finite piecewise cubic Hermite interpolating polynomial (PCHIP) family $\mathcal{U}$, with a preset 80-ms displacement bound. Temporal alignment estimates this path.

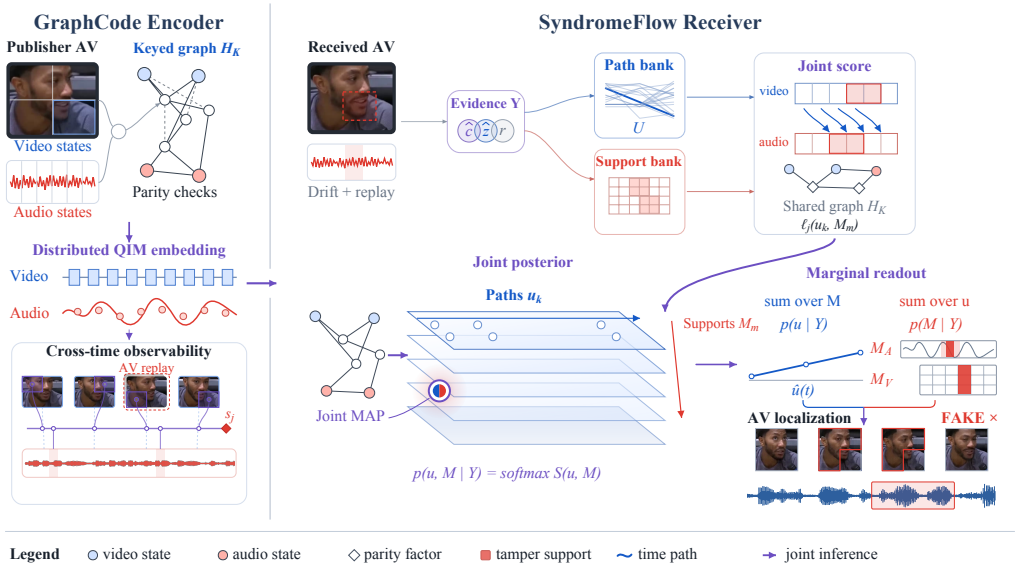

Figure 2: SyndromeFlow framework. GraphCode binds distant AV states through a keyed parity graph and distributes its factors with QIM. The receiver jointly scores temporal paths and sparse supports, then marginalizes their posterior for alignment and localization. The bottom-right example shows AV replay under affine drift; red outlines mark predicted video regions and the audio interval at threshold 0.5. The intermediate plots illustrate the inference flow.

Let the tamper support $M = (M_V, M_A)$ denote the edited video regions and audio intervals. Their binary content-state changes satisfy

$$\mathrm{supp}_g(\Delta z) \subseteq M, \tag{1}$$

since an edit can preserve a quantized state. We seek $p(M \mid Y)$, integrating the temporal path u, state changes Δz, and carrier discrepancy e. The path marginal and its entropy summarize alignment. The central challenge is separability: a smooth path should explain a global clock change, while a compact editing region should explain residual constraints that remain inconsistent under that path.

3.2 Binding distant AV content into one graph

A local watermark can reveal that one carrier changed, yet a marked donor may bring a valid watermark with it. Our GraphCode encoder binds several content states to every embedded symbol. Let z^{pub} collect binary states from video regions and 250-ms audio bins. A frozen PCA/Gray-code quantizer maps temporal-mean video thumbnails and low-band audio log powers to 16-bit states (Appendix A.2). A secret key samples a sparse heterogeneous incidence matrix H_K. For factor j, the publisher forms

$$c_j = b_j \oplus [H_K z^{\mathrm{pub}}]_j, \tag{2}$$

where b_j is a keyed code bit bound to the expected source identity. The corresponding symbol is embedded in an audio or visual carrier. After a local marked region is copied, its carrier may remain intact while its parity checks disagree with the surrounding content.

The receiver uses the received AV, the expected identity and keyed graph, and frozen channel calibration. The embedded symbols carry parity measurements of the publisher state. Candidate path u aligns the received audio before the same state mapping yields $\hat{z}(u)$. The observed factor obeys

$$\hat{c}_j = b_j \oplus \left[H_K\left(\hat{z}(u) \oplus \Delta z\right)\right]_j \oplus e_j, \tag{3}$$

Here e_j is an effective binary discrepancy from channel noise or copied carrier symbols; missing observations are masked. Deformation and content changes both alter which parity constraints agree with the received media.

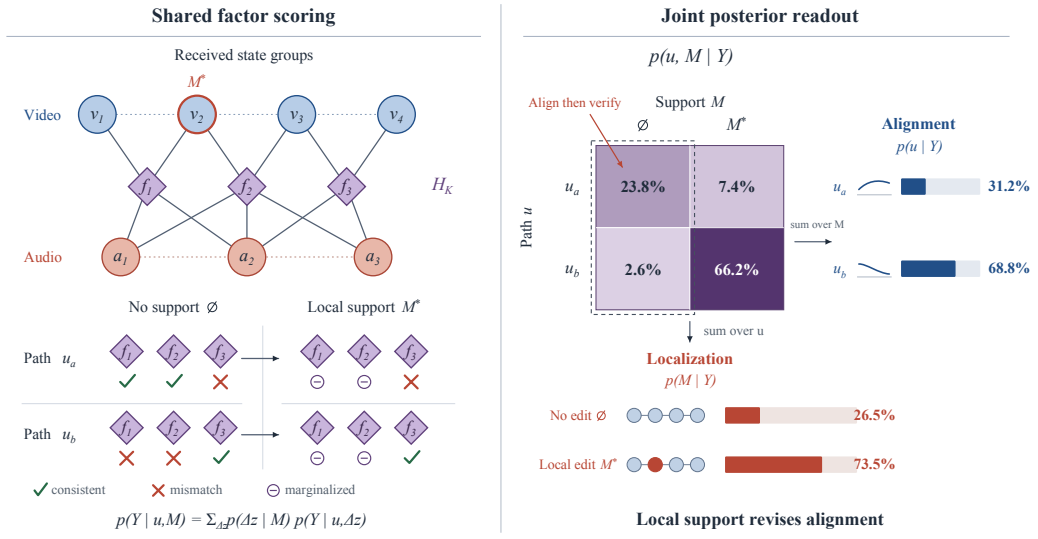

Figure 3: Local support revises temporal alignment. Nodes denote state groups. Under M^*, uncertain changes are summed out, giving equal probability to either outcome of the first two checks. The third check favors u_b. Normalizing all path–support scores yields the two marginals. Probabilities and path profiles illustrate the model.

3.3 A SHARED POSTERIOR FOR PATH AND SUPPORT

The model couples a sparse editing-region prior with conditional state changes and carrier discrepancies:

$$p(u, M, \Delta z, e \mid Y) \propto p(u)\, p(M)\, p(\Delta z \mid M)\, p(e \mid r, u, M) \times \prod_{j \in \mathcal{J}(u)} p_j(\widehat{c}_j \mid u, \Delta z, e_j, r_j, H_K). \tag{4}$$

Here $\mathcal{J}(u)$ contains available factors, and r_j collects factor j's received validity masks and QIM reliability measurements used with frozen channel parameters (Appendix A.3); $r = \{r_j\}_j$. Given e_j, p_j is the binary consistency indicator in Equation 3. The discrepancy model depends on M because editing can also replace carrier symbols; $p(\Delta z \mid M)$ permits state changes only within M. Marginalization yields

$$p(M \mid Y) = \sum_{u, e, \Delta z : \mathrm{supp}_g(\Delta z) \subseteq M} p(u, M, \Delta z, e \mid Y). \tag{5}$$

The path marginal sums over M, Δz, and e. A compact editing region that explains inconsistent factors raises the probability of compatible paths. Retaining path uncertainty until readout allows local edit evidence to refine alignment.

For comparison, the null-conditioned align–then–verify control A selects

$$\widehat{u}_{\mathrm{hard}} = \arg\max_{u \in \mathcal{U}} p(u, M = \emptyset \mid Y) \tag{6}$$

and computes support probabilities conditional on this path. It shares the same graph, observations, proposals, likelihoods, and priors with SyndromeFlow. The comparison thus isolates the effect of placing support inside the path posterior.

Figure 3 illustrates this interaction. Marginalizing uncertain changes inside a local region reduces the influence of its incident checks. The remaining checks favor u_b, so summing over support hypotheses changes the preferred path relative to selection conditioned on $M = \emptyset$. Section 4.3 measures this path-revision effect and the subsequent marginalization gain on physical AV media.

3.4 FINITE PHYSICAL RECEIVER

The physical receiver evaluates a finite approximation to Equation 4. An additive audio pilot embedded alongside QIM supplies timing evidence independently of semantic correspondence (Appendix A.2); its ambiguity is retained in the path posterior. A six-second path is a monotone PCHIP displacement with controls at 0, 3, and 6 seconds. We score all $17^3 = 4{,}913$ controls on a 10-ms grid using this keyed full-horizon audio pilot,

$$s_{\text{pilot}}(u) = \frac{1}{|I_u|} \sum_{t \in I_u} Y_A(t) p_K(t + u(t)), \tag{7}$$

and pass the top 20 paths to the parity receiver. The pilot stage reads neither the localization score nor the attack label.

A frozen convolutional damage head maps watermark-decoder responses to edit probabilities; pooling over 250-ms bins and quadrants yields p_g (Appendix A.3). Clipped log odds λ_g define size-stratified candidates:

$$\mathcal{K} = \{1, 2, 3, 4, 5, 6, 8, 10, 12, 16, 24, 32, 48, 64, 80, 96\}, \qquad M_V^{(k)} = \text{TopK}(\lambda; k). \tag{8}$$

The 23 audio intervals are all adjacent pairs among the 24 state bins. Crossing them with the 16 visual candidate masks, their single-modal counterparts, and the null state gives $16 \times 23 + 16 + 23 + 1 = 408$ hypotheses. The null, audio-only, video-only, and joint-AV families each receive prior mass $1/4$, shared uniformly within each family. We normalize $20 \times 408 = 8{,}160$ path–region scores per record, including moving and disconnected masks.

For each path, the receiver forms $s(u) = \widehat{c} \oplus b \oplus H_K \widehat{z}(u)$. A candidate region specifies where the state may have changed. We assign independent $\text{Bernoulli}(\rho)$ changes to its bits and fix changes outside it to zero. If $m_j(M)$ counts eligible bits incident to factor j, marginalizing these changes gives

$$q_j(M) = \Pr([H_K \Delta z]_j = 1 \mid M) = \frac{1 - (1 - 2\rho)^{m_j(M)}}{2}. \tag{9}$$

The physical graph uses distinct bit variables across factors, making this marginal factorization exact conditional on a candidate region. Groups can participate in several factors through different bits. We use $\rho = 1/2$: a touched factor admits either parity outcome, while untouched factors retain their alignment evidence.

This gives each support hypothesis a direct role in alignment. A candidate region can explain failed checks through uncertain content changes, while checks unconnected to it continue to test temporal correspondence. A path compatible with these remaining checks can therefore gain posterior weight when the region is considered. Summing over candidate regions carries this local explanation into the global path distribution.

Combining $q_j(M)$ with calibrated channel errors gives the predicted carrier-bit distribution. Video carriers contribute extracted symbols; audio inference sums unknown carrier words against their QIM likelihood. Appendix A.3 defines these terms and the normalized score $q_{\text{fin}}(u, M \mid Y)$, whose marginals produce the localization maps.

Carrier realization. We divide each host into 24 quarter-second state bins and four visual regions. The keyed graph contains 32 parity factors; each touches four separated video states and two audio states. Eighteen factors are carried by audio QIM and 14 by video QIM. A 32-bit application message is protected by a distributed $[46, 32]$ visual codeword across the clip. The audio branch uses activity-aware differential subband QIM. The visual branch calibrates embedding strength per host to a 37-dB luma target. These choices create the state and reliability measurements required by Equation 4; the inference model remains compatible with stronger carrier implementations.

3.5 SCALABLE GROUP-SPARSE INFERENCE

For longer sequences, the controlled graph study specializes the model to nonzero changes in each active group and uses belief propagation:

$$M_t \sim \text{Bernoulli}(s_0/T), \qquad p(\Delta z_{g_t} = x \mid M_t = 1) = \frac{1}{2^{|g_t|} - 1}, \; x \neq 0. \tag{10}$$

Here T counts temporal groups, and s_0 is their expected active count, held fixed across sequence lengths. For each path, the observed syndrome

$$s_j(u) = \widehat{c}_j \oplus b_j \oplus [H_K \widehat{z}(u)]_j \tag{11}$$

defines a noisy XOR factor. Damped sum-product produces path-conditional bit marginals and a Bethe evidence $\widehat{Z}_u$. We then couple paths through

$$q(u \mid Y) = \frac{p(u)\widehat{Z}_u}{\sum_{u'} p(u')\widehat{Z}_{u'}}, \qquad q(\Delta z_i = 1 \mid Y) = \sum_u q(u \mid Y) q_u(\Delta z_i = 1). \tag{12}$$

Sparse XOR updates use prefix and suffix products, giving $O(PEI)$ time for P paths, E graph edges, and I iterations, with $O(E)$ memory when paths are streamed. Entropy of $q(u \mid Y)$ summarizes the remaining alignment uncertainty.

4 Experiments and Discussion

We first compare complete watermark systems, then separate path revision from marginalization on physical AV media and examine their relation to alignment uncertainty. Appendix A details the receiver and its calibration.

4.1 Experimental protocol

Each six-second host contains 150 frames at 224×224, 25 fps, and 16-kHz audio. We use the graph and distributed $[46, 32]$ codeword described in Section 3. Audio QIM uses step 8×10^{-4}; visual strength is calibrated to a 37-dB luma target. OpenH264 at approximately 760 kbit/s and AAC at 128 kbit/s yield 36.38-dB mean luma PSNR, 0.0671% QIM BER, and 96.32% reliable-block retention.

For each received clip, the receiver reduces 4,913 PCHIP paths to 20 pilot proposals and pairs them with 408 support candidates. The pilot temperature is 10^{-5}, the audio likelihood has unit temperature, and the localization threshold is 0.5. Candidate construction and visual-score calibration are frozen on ten development sources; all matched readouts use identical observations and calibration. Paired percentile intervals resample sources, with 1,000 replicates for the physical confirmation; controlled graph studies use 10,000 replicates over their independent graph units.

For the scalable receiver, we run damped loopy sum-product on three candidate paths and time it in a single CPU process. Appendix A gives the graph construction and frozen thresholds.

Datasets and splits. We use authentic media from the research-only AV-Deepfake1M++ release (Cai et al., 2025). Selection follows corpus metadata, decodability, and deterministic media hashes. The panel contains 150 source-disjoint hosts: 100 VoxCeleb2 identities and 50 LRS3 videos, with one host per identity or source video. Ten of these hosts are used for receiver development and calibration; the remaining 140 form the source-disjoint confirmation. The underlying video evidence is reused only within the corresponding source and attack condition when an experiment changes the audio deformation. Dataset media are never redistributed.

The main comparisons use the 140 confirmation sources. Six benign paths per source and four attack backgrounds (± 60-ms offsets and rising/falling affine drift), crossed with marked donor/replay and four support geometries, give 840 benign samples and 4,480 AV attack records per method. The four geometries include a short quadrant, an L shape, a moving diagonal, and a split burst; all were absent from proposal development. Video copying replaces luma within each support and retains recipient chroma. All methods share source and donor identities, attack supports, timing backgrounds, and the H.264 channel. These records combine independently processed modalities: repeated audio backgrounds reuse the same attacked video, and repeated visual shapes reuse the same attacked audio. Confidence intervals therefore resample source videos. Controlled graph experiments use 32 fresh graph seeds and 6,144 synthetic records.

4.2 Main comparison: complete watermark systems

Table 1 compares five published visual watermark systems with SyndromeFlow under the same four-shape protocol. All use released weights and retain their embedding strengths and score functions.

Table 1: Visual system comparison on 4,480 marked-copy attacks from 140 speaking-person sources. All methods share the four-shape Y-copy protocol. Intervals use 1,000 source-cluster bootstrap replicates. IoU is pooled over video groups; FLR is the benign false-localization rate. External methods evaluate one identity video per source; SyndromeFlow evaluates six benign audio backgrounds.

Method	Video AUPRC ↑	95% CI	Hard IoU ↑	FLR (%) ↓
EditGuard (Zhang et al., 2024a)	.0862	[.0713,.1197]	.0000	.000
WAM (Sander et al., 2025)	.0594	[.0568,.0628]	.0193	2.984
MaskWM-D (Hu et al., 2025)	.0621	[.0594,.0652]	.0567	39.561
OmniGuard (Zhang et al., 2025b)	.0603	[.0581,.0629]	.0252	3.899
ReImage (Kim & Seo, 2026)	.0705	[.0645,.0785]	.0439	5.580
SyndromeFlow	**.9937**	[.9906,.9959]	**.9228**	.017

Table 2: Audio system comparison on the same 140-source confirmation panel: 4,480 AV attack records and 840 benign records per method. Each system uses its own watermark; AudioSeal uses its native presence score.

Method	Audio AUPRC ↑	Hard IoU ↑	FLR (%) ↓
AudioSeal (San Roman et al., 2024)	.0718	.0507	66.409
SyndromeFlow	**.6371**	**.5187**	**.362**

EditGuard, WAM, and MaskWM-D repeat source messages across frames without explicit frame-position binding; OmniGuard uses its released localization branch, and ReImage embeds shuffled image content. SyndromeFlow distributes a time-bound code over the AV clip. This comparison evaluates localization when copied content carries its watermark. Payloads and distortion levels remain method-specific; Appendix B.4 records the inference adapters. Table 5 complements this comparison with a carrier-level diagnostic at matched RGB-PSNR targets.

Table 2 shows a similar advantage over AudioSeal's watermark-presence score. Copying can preserve presence, while benign timing changes can disturb extraction. Positive audio bins occupy 8.33% of the timeline, providing a random-ranking reference.

4.3 Mechanism decomposition on physical AV

Table 3 separates two operations using one finite posterior, denoted by $q(u, M \mid Y)$. Readout A conditions support inference on the null-region path $\widehat{u}_{\mathrm{hard}}$ from Equation 6. Readout B uses the support-aware MAP path $u_* = \arg\max_u \sum_M q(u, M \mid Y)$ and reports $q(M \mid u_*, Y)$. Readout C averages over paths to obtain $\sum_u q(u, M \mid Y)$. A→B measures support-driven path revision; B→C measures residual-path marginalization. B and C report the same MAP path.

Two pilot controls select the path by correlation or pilot likelihood plus the frozen path prior, then evaluate the same conditional support posterior $q(M \mid u, Y)$.

The pilot-plus-prior control achieves 0.5861 audio AUPRC; full SyndromeFlow raises it to 0.6371 and reduces path MAE from 2.184 to 0.949 ms. In the A→B→C decomposition, support-driven path revision supplies most of the improvement: audio AUPRC rises by 0.0661 from A to B. Marginalization adds 0.0047 AUPRC and reduces audio Brier error.

4.4 When does coupling matter?

We group attacks by normalized pilot-path entropy, computed before support inference. The four equal-sized groups contain 1,120 attacks each; their boundaries depend only on pilot evidence. Table 4 relates the two readout effects to this initial uncertainty.

Path correction reduces MAE by 4.67 ms more in Q4 than in Q1. Marginalization adds 0.0134 audio AUPRC in Q4, where the fused posterior retains an average 4.21% probability outside its MAP path. The corresponding mass in Q1 is 0.004%. This separates initial correspondence ambiguity from

Table 3: Physical readout comparison on 4,480 attacks and 840 benign records from 140 source-disjoint hosts. All rows share the watermark, observations, proposals, support priors, and localization likelihoods and thresholds; path selection and readout vary. C is full SyndromeFlow. Brier is mean squared audio probability error; A-FLR and V-FLR are benign false-localization rates.

Readout	Audio AUPRC ↑	Video AUPRC ↑	Path MAE ↓	Brier ↓	A-FLR (%)	V-FLR (%)
Pilot Top-1	.5808	.9935	2.734 ms	.04640	.441	.017
Pilot + prior	.5861	.9936	2.184 ms	.04596	.461	.020
A: Null path	.5663	**.9943**	5.625 ms	.04960	.352	.017
B: Support MAP	.6324	.9936	**0.949 ms**	.04258	.372	.017
C: Path marginal	**.6371**	.9937	**0.949 ms**	**.04247**	.362	.017

Table 4: Physical mechanism effects across pilot-entropy quartiles. Path correction is the mean MAE reduction from A to B; marginal gain is the pooled audio AUPRC change from B to C. Higher values indicate larger improvements.

Pilot entropy	Path correction (ms) ↑	Marginal AUPRC gain ↑
Q1 (lowest)	1.79	−.00016
Q2	3.90	−.00003
Q3	6.55	+.00249
Q4 (highest)	6.46	+.01341

uncertainty remaining after joint inference. Controlled graphs provide complementary evidence: sparse or erased pilots increase the value of path marginalization (Appendix C.2).

This interaction has a direct probabilistic interpretation. Let $s_g(u) = q(g \in M \mid u, Y)$ be the tamper probability for group g under path u. The conditional and marginal readouts satisfy

$$s_C(g) - s_B(g) = \sum_{u \neq u_*} q(u \mid Y)\big[s_g(u) - s_g(u_*)\big]. \tag{13}$$

Each alternative path contributes according to its posterior weight and its disagreement with the MAP path's localization. Path revision selects the leading temporal explanation; marginalization carries the remaining disagreement into localization probabilities.

4.5 Discussion

Joint inference maintains a shared distribution over temporal alignment and tamper support, and the decomposition reveals two distinct effects. Support-driven path revision is dominant in the matched physical comparison: local edit evidence can change the preferred alignment, reducing alignment-induced authentication residuals. Marginalization has a complementary probabilistic role. When support evidence concentrates the posterior on one path, MAP conditioning and path averaging yield nearly identical localization probabilities. When residual path uncertainty remains, averaging can improve pooled localization and probability scores.

Controlled studies clarify how this coupling operates under ambiguous evidence. Group sparsity improves support recovery (Appendix C.3), while weak pilots increase the benefit of path averaging (Appendix C.2). Belief propagation extends inference to substantially longer controlled sequences (Appendix C.6). Together, these results identify the central contribution as coupling alignment and edit hypotheses through shared authentication evidence, with path revision and marginalization serving complementary roles.

5 Conclusion

SyndromeFlow places temporal deformation and sparse AV recomposition in one watermark-authentication posterior. Physical decomposition shows that support-driven path revision supplies the main alignment gain, while marginalization adds a smaller pooled localization gain concentrated in the highest pilot-entropy group. Future work can extend the physical receiver to continuous paths and finer audio localization.

REPRODUCIBILITY STATEMENT

The submission package records all frozen path grids, proposal cardinalities, carrier parameters, priors, thresholds, and source-level split hashes. Table values are derived from machine-readable results using documented aggregation procedures. Statistical resampling uses the source as the cluster and stores its random seed. The appendix provides the attack definitions, likelihood factorization, proposal coverage, and controlled-graph studies. We plan to release the receiver code, experiment configurations, aggregation scripts, and derived metadata. Source media remain subject to the licenses of their original datasets.

ETHICS STATEMENT

This work studies provenance signals for localizing edits in published media. Experiments use licensed research datasets and preserve their distribution conditions. The released implementation will operate on researcher-published watermarks and will exclude source media. Watermark localization can also be used to study or adapt editing strategies; documenting the threat model, carrier assumptions, and evaluation protocols supports responsible comparison.

AI USE STATEMENT

Generative AI tools assisted with polish writing and search for literature and validate the novelty of ideas. The research team reviewed scientific hypotheses, protocol decisions, data selection, reported values, and citations against source material. The authors retain responsibility for the manuscript and its claims.

REFERENCES

Adnan Alattar, Ravi Sharma, and John Scriven. A system for mitigating the problem of deepfake news videos using watermarking. *Electronic Imaging*, 32(4):117–1–117–10, 2020. doi: 10.2352/ISSN.2470-1173.2020.4.MWSF-117.

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. AV-Deepfake1M++: A large-scale audio-visual deepfake benchmark with real-world perturbations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13686–13691, 2025. doi: 10.1145/3746027.3761979.

Brian Chen and Gregory W. Wornell. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE Transactions on Information Theory*, 47(4):1423–1443, 2001. doi: 10.1109/18.923725.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-View Lip-Reading, Asian Conference on Computer Vision*, 2016. URL https://robots.ox.ac.uk/~vgg/software/lipsync/.

Marco Cuturi and Mathieu Blondel. Soft-DTW: A differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903, 2017. URL https://proceedings.mlr.press/v70/cuturi17a.html.

Mikita Dvornik, Isma Hadji, Konstantinos G. Derpanis, Animesh Garg, and Allan D. Jepson. Drop-DTW: Aligning common signal between sequences while dropping outliers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13782–13793, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/729c68884bd359ade15d5f163166738a-Abstract.html.

Clara Fernandez-Labrador, Mertcan Akçay, Eitan Abecassis, Joan Massich, and Christopher Schroers. DiVAS: Video and audio synchronization with dynamic frame rates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26846–26854, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/

Fernandez-Labrador_DiVAS_Video_and_Audio_Synchronization_with_Dynamic_Frame_Rates_CVPR_2024_paper.html.

Runyi Hu, Jie M. Zhang, Shiqian Zhao, Nils Lukas, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Mask image watermarking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 146313–146346, 2025. doi: 10.52202/085713-4895.

Minyoung Kim, Sehwan Park, Sungmin Cha, and Paul Hongsuck Seo. Cross-modal watermarking for authentic audio recovery and tamper localization in synthesized audiovisual forgeries. In *Interspeech 2025*, pp. 5103–5107, 2025. doi: 10.21437/Interspeech.2025-316.

Minyoung Kim and Paul Hongsuck Seo. Robust image self-recovery against tampering using watermark generation with pixel shuffling. In *CVPR Findings*, pp. 8877–8886, 2026. URL https://eurominyoung186.github.io/ReImage/.

Pascal Lefèvre, Philippe Carré, Caroline Fontaine, Philippe Gaborit, and Jiwu Huang. Efficient image tampering localization using semi-fragile watermarking and error control codes. *Signal Processing*, 190:108342, 2022. doi: 10.1016/j.sigpro.2021.108342.

Chang-Tsun Li and Yinyin Yuan. Digital watermarking scheme exploiting nondeterministic dependence for image authentication. *Optical Engineering*, 45(12):127001, 2006. doi: 10.1117/1.2402932.

Yao-Chung Lin, David Varodayan, and Bernd Girod. Image authentication and tampering localization using distributed source coding. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pp. 393–396, 2007. doi: 10.1109/MMSP.2007.4412899.

Huajian Liu and Martin Steinebach. Semi-fragile watermarking for image authentication with high tampering localization capability. In *Proceedings of the Second International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution*, pp. 143–152, 2006. doi: 10.1109/AXMEDIS.2006.48.

Jiale Meng, Runyi Hu, Jie Zhang, Zheming Lu, Ivor W. Tsang, and Tianwei Zhang. Unifying watermarking via dimension-aware mapping. *arXiv preprint arXiv:2602.03373*, 2026. URL https://arxiv.org/abs/2602.03373.

Ronaldo Rigoni, Pedro Garcia Freitas, and Mylène C. Q. Farias. Detecting tampering in audio-visual content using QIM watermarking. *Information Sciences*, 328:127–143, 2016. doi: 10.1016/j.ins.2015.08.040.

Robin San Roman, Pierre Fernandez, Hady Elsahar, Alexandre Défossez, Teddy Furon, and Tuan Tran. Proactive detection of voice cloning with localized watermarking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 43180–43196, 2024. URL https://proceedings.mlr.press/v235/san-roman24a.html.

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IkZVDzdC8M.

Luyang Si, Leyi Pan, and Lijie Wen. mAVE: A watermark for joint audio-visual generation models. *arXiv preprint arXiv:2603.07090*, 2026. URL https://arxiv.org/abs/2603.07090.

Jianren Wang, Zhaoyuan Fang, and Hang Zhao. AlignNet: A unifying approach to audio-visual alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3309–3317, 2020. URL https://openaccess.thecvf.com/content_WACV_2020/html/Wang_AlignNet_A_Unifying_Approach_to_Audio-Visual_Alignment_WACV_2020_paper.html.

Bokang Zeng, Zheng Gao, Xiaoyu Li, Xiaoyan Feng, and Jiaojiao Jiang. LAVA: Layered audio-visual anti-tampering watermarking for robust deepfake detection and localization. *arXiv preprint arXiv:2604.23957*, 2026. URL https://arxiv.org/abs/2604.23957.

Chuan Zhang, Zihan Li, Zihao Xu, Xuhao Ren, and Liehuang Zhu. SepVAMark: Deep separable visual-audio fusion watermarking for source tracing and deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8910–8919, 2025. doi: 10.1145/3746027.3755783.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11964–11974, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_EditGuard_Versatile_Image_Watermarking_for_Tamper_Localization_and_Copyright_Protection_CVPR_2024_paper.html.

Xuanyu Zhang, Youmin Xu, Runyi Li, Jiwen Yu, Weiqi Li, Zhipei Xu, and Jian Zhang. V^2A-Mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9818–9827, 2024b. doi: 10.1145/3664647.3680904.

Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3008–3018, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_OmniGuard_Hybrid_Manipulation_Localization_via_Augmented_Versatile_Deep_Image_Watermarking_CVPR_2025_paper.html.