Beyond Isolated Clips: Structured Learning for Temporal Forgery Localization

Anonymous authors
Paper under double-blind review

Abstract

Temporal forgery localization must identify the extent of a manipulated interval, not only assign suspicious scores to short clips. We study speech-assisted visual localization in talking-person video, where the input is one audio-video clip and the output is a set of visual manipulation intervals on its original timeline. Our structured localizer scores a common graph of 25-Hz candidate segments, trains the scores through a globally normalized distribution over complete temporal labelings, and uses an input-derived speech prior only at inference. This separates the learning objective from the choice of speech-guided readout. On a locked 212-video ArEnAV preview, the resulting U+S system reaches 46.59/33.76/2.65 AP at IoU 0.50/0.75/0.95 for Diff2Lip-trained checkpoints and 34.03/23.91/2.62 for LatentSync-trained checkpoints. A source-calibrated independent model given the same decoder reaches 28.42/15.39/0.10 and 26.87/6.55/0.08, respectively; the AP75 gains are 18.37 and 17.36 percentage points. Cross-generator direction rows and a strict visual LAV-DF transfer evaluation provide complementary evidence, while a matched U-max control exposes a segment-versus-frame ranking trade-off. Together, these results support globally normalized segment learning as a practical complement to independent clip scoring under the fixed protocol, with each evaluation scope reported separately.

1 Introduction

A collection of plausible clip scores is not yet a temporal localization. Several overlapping clips may contain evidence of manipulation while disagreeing about where the modification begins and ends. For a user inspecting a talking-person video, the useful output is not a list of suspicious clips but intervals that account for the altered visual content. We study speech-assisted visual forgery localization: a model receives one audio-video clip and returns visual manipulation intervals without an original recording, editing metadata, or a parent reference at deployment. This setting complements recent audio-visual forgery datasets and systems that evaluate temporal boundaries directly (Cai et al., 2022; 2023; 2024).

A common pipeline defines temporal units, classifies them, and combines their scores. Fixed windows provide coverage, while speech-derived word intervals offer a useful time organization for talking-person video. These design choices do not determine the learning objective. Independent segment classification can learn useful local evidence, but it does not directly train competing segments to form a complete authenticity labeling of the timeline. Temporal action localization has long separated segment scoring from sequence-level organization (Zhang et al., 2022), and structured sequence models make the distinction explicit (Sarawagi & Cohen, 2004; Kong et al., 2016).

We develop a structured segment localizer, which we call U, that puts this competition inside training. The model assigns visual compatibility scores to candidate segments and evaluates complete legal segment-label paths in a globally normalized distribution. Dense temporal supervision fixes the authenticity state of each time cell, while same-state subdivisions remain latent. The loss therefore compares all legal explanations with the explanations compatible with the observed labeling. The resulting gradient increases evidence for paths that explain

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.

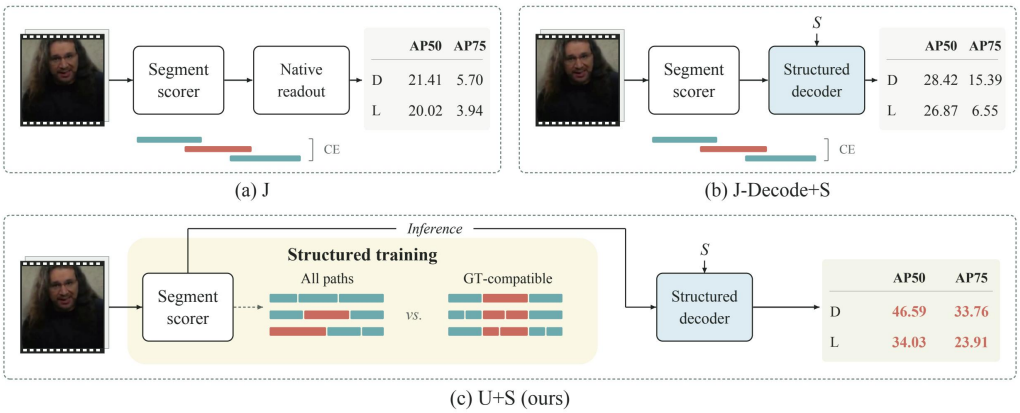


Figure 1: Controlled comparison behind Table 1. (a) J uses independent segment learning and native readout; (b) J receives the common structured decoder and speech prior S; (c) $U + S$ trains all legal paths against ground-truth-compatible paths and uses S only for inference. D and L denote Diff2Lip and LatentSync training directions.

the complete sequence and decreases evidence for alternatives, rather than treating every candidate as an independent example.

The formulation separates visual segment learning from speech-guided inference. We train with a neutral duration prior and use a designed prior derived from current-input Whisper word-point timestamps during inference. The timestamps guide candidate weighting but do not define an immutable partition, and they are not treated as a calibrated boundary posterior. With unavailable or unusable words, the system falls back to the neutral prior. The final readout uses fake-state frame marginals, fixed thresholds, connected runs, coordinate deduplication, and the released SoftNMS-style postprocessing. All speech information comes from the clip being examined.

Our controlled comparison keeps the visual VideoMAE/LoRA computation, candidate graph, speech readout, and decoder fixed while changing the learning objective. The independent expected-CE learner is denoted by J; J receives a source-only affine calibration and the common structured decoder in the matched J-Decode+S row. This isolates whether structured training adds value when candidate support and speech guidance are already shared. On the locked preview, Diff2Lip-trained $U + S$ improves AP75 by 18.37 points over matched J-Decode+S, and LatentSync-trained $U + S$ improves it by 17.36 points. The same comparison also improves AP50, AP95, frame AP, and AR50 for both training configurations. Figure 1 shows the three stages directly: native independent readout, post-hoc structured decoding with S, and structured training with the same inference decoder.

We report three complementary scopes: the locked 212-video preview, corrected train-to-evaluation directions, and a strict visual LAV-DF transfer diagnostic. Each table labels its split and provenance explicitly. On LAV-DF, frozen local-input U rows are paired with the released external checkpoint as a scope-specific context. A matched U-max experiment further characterizes how compatible-path aggregation changes segment and frame ranking.

This paper makes three concrete contributions. First, it gives a globally normalized segment-learning formulation for speech-assisted visual forgery localization with explicit compatible-path supervision. Second, it isolates the contribution of structured training from speech-guided decoding using a shared decoder and source-only calibration, obtaining the AP75 gains above on both generator-specific training configurations. Third, it packages corrected direction, transfer, ablation, degradation, and public-scope evidence so that readers can interpret the result at the scope supported by the data.

2 Related Work

Visual and audiovisual forgery localization. LAV-DF introduced a content-driven audio-visual dataset and temporal localization baselines (Cai et al., 2022; 2023). AV-Deepfake1M expanded the scale and made the benchmark useful for multimodal temporal localization (Cai et al., 2024). ArEnAV targets audiovisual forgery localization with diverse speech and generation conditions (Kuckreja et al., 2026); the present study uses its locked preview and retains its visual-localization scope. Recent systems such as WAFL and AuViRe combine speech or audio-visual representations with temporal prediction (Wang et al., 2026; Koutlis & Papadopoulos, 2026). We place these systems in a scope-separated context table so their input contracts remain explicit.

Segment-level learning and structured prediction. ActionFormer demonstrates the value of transformer-based temporal action localization with direct segment prediction (Zhang et al., 2022). Semi-Markov conditional random fields provide a classical formulation for segmental sequence structure (Sarawagi & Cohen, 2004); segmental recurrent models learn segment representations while summing over latent boundaries (Kong et al., 2016). Our contribution is empirical and task-specific: we use a fixed candidate graph and ask whether globally normalized compatible-path training improves the localization readout over independent segment learning under the same downstream decoder. We adopt standard dynamic programming so that the comparison remains focused on the learning objective.

Speech-derived timing and visual encoders. Whisper supplies current-input speech timestamps (Radford et al., 2023); these timestamps are used here as point estimates for a designed prior over candidate timing. VideoMAE provides the self-supervised visual encoder (Tong et al., 2022), while LoRA supplies the low-rank adaptation mechanism (Hu et al., 2022). These components provide a shared engineering substrate; the scientific comparison concerns how candidate segment evidence is trained and read out.

3 Method

3.1 Task and candidate segment graph

Let X denote the visual frame sequence and A the synchronized audio track. We use a 25-Hz grid with valid frame cells $0, \dots, T-1$ and binary visual-authenticity labels $y_{0:T-1}$. A candidate edge is a half-open interval $e = (s, t, z)$ with integer grid endpoints $0 \le s < t \le T$, state $z \in \{0, 1\}$, and $1 \le t - s \le 50$ frames. A legal path π starts at frame 0, ends at frame T, and concatenates adjacent edges. The final output is obtained from fake-state frame marginals and fixed threshold/run postprocessing; it is not restricted to a single MAP path.

The shared visual frontend is a frozen VideoMAE-base prefix for blocks 0–7 followed by rank-4 Q/V LoRA in blocks 8–11. Full RGB input is resized to 224 pixels, processed in 16-frame windows at 25 Hz with stride 8, and combined on the valid two-frame supports. A common segment head applies layer normalization to the 3072-dimensional token, a 128-dimensional projection with GELU, two residual dilated temporal convolutions, and a 513-to-128-to-1 candidate head. The candidate head receives mean, start, end, difference, and log-length features from the shared temporal tokens. Repeating outputs to the 25-Hz coordinate grid changes the coordinate representation, not the visual observations.

3.2 Visual evidence and complete temporal labelings

For each edge e, the head produces a fake logit $r_\theta(X, e)$. Its edge-state potential is

$$\Psi_\theta(X, e, z) = \begin{cases} (t - s) \log \sigma(r_\theta(X, e)), & z = 1, \\ (t - s) \log \sigma(-r_\theta(X, e)), & z = 0. \end{cases}$$

The path score is additive, $\Psi_\theta(X, \pi) = \sum_{e \in \pi} \Psi_\theta(X, e, z_e)$. This candidate-dependent pooling means that changing an edge changes the segment statistics while the VideoMAE window encoder remains shared.

Let $N(\pi)$ be the neutral duration prior. For every start node it normalizes $\exp(-(t-s-1)/8)$ over the outgoing edges, so it is a local length prior rather than a globally uniform path weight. The all-path partition and compatible-path partition are

$$Z_N(X) = \sum_{\pi} N(\pi) \exp\{\Psi_\theta(X, \pi)\}, \qquad Z_N(X, y) = \sum_{\pi \in \mathcal{C}(y)} N(\pi) \exp\{\Psi_\theta(X, \pi)\}.$$

Here $\mathcal{C}(y)$ contains paths whose every edge is wholly inside one constant-state run, namely $y_s = \cdots = y_{t-1}$. Authenticity transitions are therefore fixed by the observed frame labels, while the subdivision of each same-state run remains latent. Because unit edges are included, every binary labeling on the valid grid has at least one compatible path. The training objective, normalized by valid frames, is

$$\mathcal{L}(\theta) = \log Z_N(X) - \log Z_N(X, y).$$

For a frozen prior, its explanatory gradient is the partition-function identity

$$\nabla \mathcal{L} = \mathbb{E}_{\text{all}}[\nabla \Psi_\theta] - \mathbb{E}_{\text{compatible}}[\nabla \Psi_\theta].$$

We use this as an interpretation of the learning signal, not as a new theorem.

3.3 Speech-guided inference

Whisper word timestamps are mapped to the current video clock and quantized to the same 25-Hz grid. Let $W(A)$ be the union of recognized word spans, expanded by five grid frames. The inference prior $S(e \mid A)$ reuses the neutral duration potential and multiplies it by a fixed endpoint factor

$$b(t; A) = 0.1 + 0.9 \exp[-d(t, A)^2/8],$$

where $d(t, A)$ is the distance in grid frames to the nearest start or end point of a recognized word span when $t \in W(A)$; outside $W(A)$ we set $b(t; A) = 1$. S is normalized per start. It is a designed input prior: text and recognition confidence do not score forgery, and it is not a calibrated posterior over editing boundaries. If no usable words are available, $S = N$.

Inference uses

$$p_S(\pi \mid X, A) \propto S(\pi \mid A) \exp\{\Psi_\theta(X, \pi)\}.$$

We compute fake-state frame marginals by inside–outside dynamic programming, threshold them at $0.1, 0.3, 0.5, 0.7, 0.9$, form connected runs, assign mean run scores, deduplicate coordinates, and apply fixed SoftNMS with 101 proposals. Ground-truth labels enter only $\mathcal{C}(y)$ during training; they are not used by inference.

3.4 Compact algorithm and cost

At training time: (1) encode valid frame windows with the shared frontend; (2) build all native-grid edges of lengths 1–50; (3) score edges and construct Ψ; (4) run the forward recurrence for Z_N and $Z_N(\cdot, y)$; (5) backpropagate their log difference. At inference: (6) transcribe only the current audio; (7) construct S; (8) run the same recurrence with S; (9) read out fake-state marginals and fixed postprocessing. The graph has $O(TL)$ edges for maximum length $L = 50$, and the recurrence has linear dependence on the number of edges. The notation fixes the boundary convention: T is the first frame after the valid support, so all intervals are half-open. The locked driver records a 261.3-second wall-clock pass for 26 readout conditions and 12 physical forward computations; this is an implementation audit rather than a hardware throughput claim.

4 Experiments

4.1 Evaluation protocol

We use the locked ArEnAV preview test: 212 videos, 106 parent groups, 106 real clips, 106 visual-only fake clips, and 106 annotated intervals. Checkpoints and threshold conventions were selected from source-only validation and then sealed before preview scoring. We report

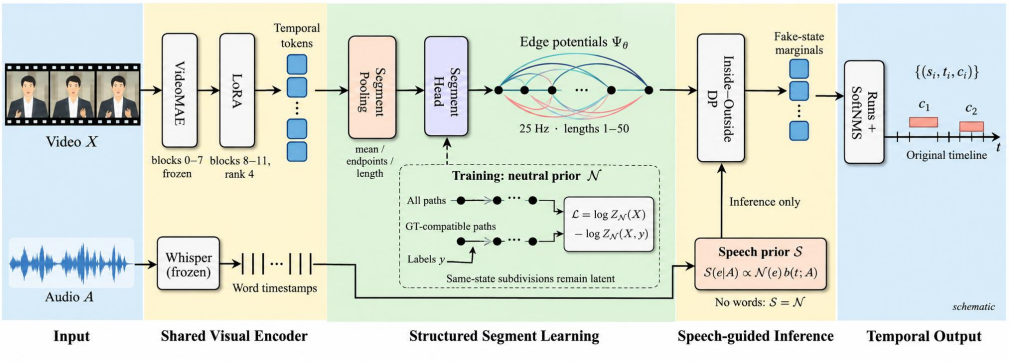


Figure 2: Method overview. The shared VideoMAE–LoRA encoder turns video and speech timing into temporal tokens and edge potentials on a 25-Hz graph with segment lengths 1–50. Neutral-prior training contrasts all legal paths with ground-truth-compatible paths; the speech prior S is injected only at inference, where inside–outside marginals become deduplicated interval proposals.

average precision (AP) at IoU 0.50, 0.75, and 0.95, framewise average precision (frame AP), and average recall at 50 proposals (AR50). All metric values are percentages. The preview is a fixed 212-video evaluation scope, and the pooled score is accompanied by its two generator-specific directions.

The primary controlled comparison uses the same VideoMAE/LoRA frontend, candidate support, speech prior, and final decoder. We denote independent expected-CE segment learning by J; it uses the same speech prior. J-Decode applies the exact common dynamic program and a source-hold-only affine calibration before the speech readout. U is the globally normalized structured localizer. The calibration uses no target labels.

4.2 Main controlled comparison

Table 1 gives pooled and direction-specific results. On the pooled preview, U+S reaches 46.59 AP50 and 33.76 AP75 for Diff2Lip-trained checkpoints, and 34.03 and 23.91 for LatentSync-trained checkpoints. Source-calibrated J-Decode+S reaches 28.42 and 15.39, and 26.87 and 6.55. The AP75 improvements are 18.37 and 17.36 points. Exact dynamic-program equivalence was checked independently: the vectorized adapter matched the historical recurrence with zero log-partition difference, sub-micro numerical marginal differences, and identical MAP sequences.

The corrected direction table separates training and evaluation generators. Diff2Lip-to-Diff2Lip gives 43.18/30.22 AP50/AP75, and Diff2Lip-to-LatentSync gives 54.61/40.65. The LatentSync-to-Diff2Lip and LatentSync-to-LatentSync rows give 34.51/24.71 and 30.73/19.50. These rows expose direction dependence and complement the pooled result; they are reported separately from the strict transfer scope.

The interval audit in Figure 3 visualizes all four corrected train-to-evaluation directions. Each panel aligns the sampled frames, the ground-truth interval, and the outputs of J, J-Decode+S, and $U+S$ on the original timeline. $U+S$ attains the highest or tied-highest best-IoU in every panel: 0.89, 0.88, 1.00, and 0.79 for Diff2Lip→Diff2Lip, Diff2Lip→LatentSync, LatentSync→Diff2Lip, and LatentSync→LatentSync, respectively.

4.3 Transfer and external context

Table 2 separates exact scopes. On strict LAV-DF visual244, the 244-video scope where the visual task is defined, frozen ArEnAV-trained local-input U obtains 88.93/73.68/1.51 AP50/AP75/AP95 for Diff2Lip-trained checkpoints and 82.41/63.63/0.64 for LatentSync-trained checkpoints. These rows quantify portability with locally derived inputs.

Table 1: Main controlled comparison on the locked ArEnAV preview ($n = 212$). AP values are percentages; the source-calibrated J-Decode row uses source-hold-only calibration and the same structured decoder as U+S.

Training direction	system	AP50	AP75	AP95	frame AP	AR50
Diff2Lip pooled	J native	21.41	5.70	0.00	34.99	38.68
Diff2Lip pooled	J-Decode+S calibrated	28.42	15.39	0.10	34.66	41.42
Diff2Lip pooled	U+S	46.59	33.76	2.65	56.22	55.75
LatentSync pooled	J native	20.02	3.94	0.01	33.36	35.75
LatentSync pooled	J-Decode+S calibrated	26.87	6.55	0.08	30.71	38.02
LatentSync pooled	U+S	34.03	23.91	2.62	59.34	42.26
D → D	U+S	43.18	30.22	1.71	55.17	49.46
D → L	U+S	54.61	40.65	2.32	58.07	70.31
L → D	U+S	34.51	24.71	1.68	56.28	38.51
L → L	U+S	30.73	19.50	3.33	64.32	50.94

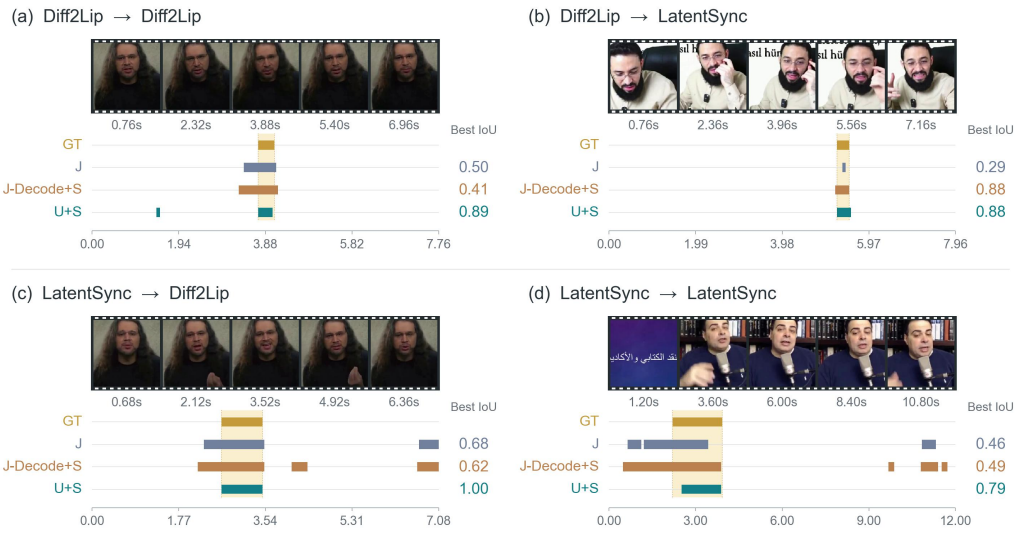


Figure 3: Interval-alignment visualization across the four corrected directions. Each panel shows sampled frames, the shaded ground-truth interval, and the temporal outputs of J, J-Decode+S, and $U + S$; the right column reports the best IoU for each system.

The released AuViRe checkpoint provides an external same-scope reference, reaching 98.81/94.74/19.42 on the same 244 IDs and evaluator. WAFL Mode B adds a 212-video input-adapter reference, with its evaluation contract recorded alongside the row.

4.4 Learning and inference ablations

The matched U-max control replaces log-sum-exp compatible-path aggregation with a max-compatible numerator and uses the same 1000-step schedule, data, seed, and common speech readout. The matched U-max control shows how compatible-path aggregation selects a different operating point: it raises segment AP50 in both generators and raises LatentSync AP75, while Table 3 records the corresponding frame-level behavior. Table 3 also contrasts source-calibrated structured decoding with structured training at the selected checkpoint.

4.5 Reading the controlled result

The pooled rows show that the gain is concentrated at the boundary-sensitive operating point. Relative to matched J-Decode+S, $U + S$ adds 18.37 AP75 points for Diff2Lip and

Table 2: Scope-separated transfer and public-system context. Values are AP50/AP75/AP95 percentages; rows are never pooled across scopes.

Scope	system	AP50	AP75	AP95	frame AP	provenance
ArEnAV common196	U+S (D)	46.88	33.43	2.50	56.44	frozen local system
ArEnAV common196	U+S (L)	32.63	22.56	2.76	58.02	frozen local system
ArEnAV common196	AuViRe	5.77	2.24	0.00	15.66	released checkpoint
LAV visual244	U (D)	88.93	73.68	1.51	90.89	local-input transfer
LAV visual244	U (L)	82.41	63.63	0.64	94.61	local-input transfer
LAV visual244	AuViRe	98.81	94.74	19.42	95.33	released native input
ArEnAV preview212	WAFL Mode B	1.58	0.36	0.00	4.43	limited adapter

Table 3: Learning and aggregation ablations. The U-max rows use a matched 1000-update schedule, seed, and common speech readout; they are separate from the principal selected step-4000 system.

Generator	system	AP50	AP75	AP95	frame AP	AR50
Diff2Lip	U frozen, 1000 step	24.54	14.07	2.00	45.50	33.96
Diff2Lip	U-max, 1000 step	27.05	14.14	2.10	36.62	39.34
LatentSync	U frozen, 1000 step	25.51	13.63	0.89	51.05	43.68
LatentSync	U-max, 1000 step	29.39	22.06	3.83	39.12	36.32
Diff2Lip	U+Decode+S source-cal control	43.36	30.81	3.11	59.04	50.85
Diff2Lip	U+S selected	46.59	33.76	2.65	56.22	55.75
LatentSync	U+Decode+S source-cal control	38.88	25.48	2.83	53.61	47.17
LatentSync	U+S selected	34.03	23.91	2.62	59.34	42.26

17.36 points for LatentSync, while also adding 21.56 and 28.63 frame-AP points. The improvement therefore appears in interval ranking and frame occupancy together, rather than only in one proposal metric. Direction rows refine this picture: the strongest AP75 row is Diff2Lip→LatentSync (40.65), while LatentSync→LatentSync is lower (19.50), making the generator-specific split visible instead of averaging it away.

The implementation audit checks that this effect is produced by the stated recurrence. The vectorized adapter and the historical recurrence have identical log partitions, sub-micro marginal differences, and identical MAP sequences. The readout then applies the same five thresholds and SoftNMS settings to every system. These checks make the numerical comparison about the training objective while keeping the decoder and postprocessing fixed.

4.6 Behavior under input degradation

Audio SNR perturbations preserve the corrected $U + S$ AP50/AP75 near the clean row, and the compression rows characterize the model's operating behavior across visual quality levels. The matched ordinary J rows provide the same reference grid. Table 4 reports representative values; the appendix gives the complete AP metrics.

Figure 4 complements the interval lanes with fake-state probability profiles. The shaded region marks the ground-truth interval, dashed and dash-dotted curves show J and J-Decode+S, and the solid teal curve shows $U + S$. Across the four panels, $U + S$ supplies a top-ranked or second-ranked proposal with IoU 0.78/1.00/0.95/0.98, making the concentration of its temporal evidence visible alongside the aggregate AP metrics.

5 Discussion

The controlled result supports a specific interpretation: learning segment evidence as parts of complete temporal explanations is useful under a fixed candidate graph and fixed speech-guided decoder. The comparison isolates the learning objective because J-Decode receives the same dynamic program and source-only calibration. The four-direction visual audit

Table 4: Behavior under input degradation for U+S and the matched ordinary J baseline. Values are AP50/AP75 percentages on $n = 212$; strong video compression causes a marked drop.

		U+S				J native			
Generator	condition	AP50	AP75	AP95	frame AP	AP50	AP75	AP95	frame AP
D	clean	46.59	33.76	2.65	56.22	19.99	6.17	0.06	29.43
D	audio SNR20	46.17	33.99	2.48	57.23	21.93	7.34	0.01	29.45
D	audio SNR10	47.20	36.27	4.63	58.56	19.21	5.16	0.00	29.11
D	video CRF23	19.27	10.82	1.88	27.86	0.76	0.23	0.00	8.01
D	video CRF35	2.61	1.40	0.32	7.12	0.25	0.11	0.00	5.65
L	clean	34.03	23.91	2.62	59.34	20.08	7.16	0.12	31.79
L	audio SNR20	35.62	25.38	2.31	59.83	19.88	5.97	0.07	30.85
L	audio SNR10	38.24	27.70	2.49	60.52	19.58	5.80	0.13	31.45
L	video CRF23	6.46	4.12	0.56	13.35	1.56	0.29	0.00	5.81
L	video CRF35	0.43	0.03	0.00	5.33	0.11	0.01	0.00	4.44

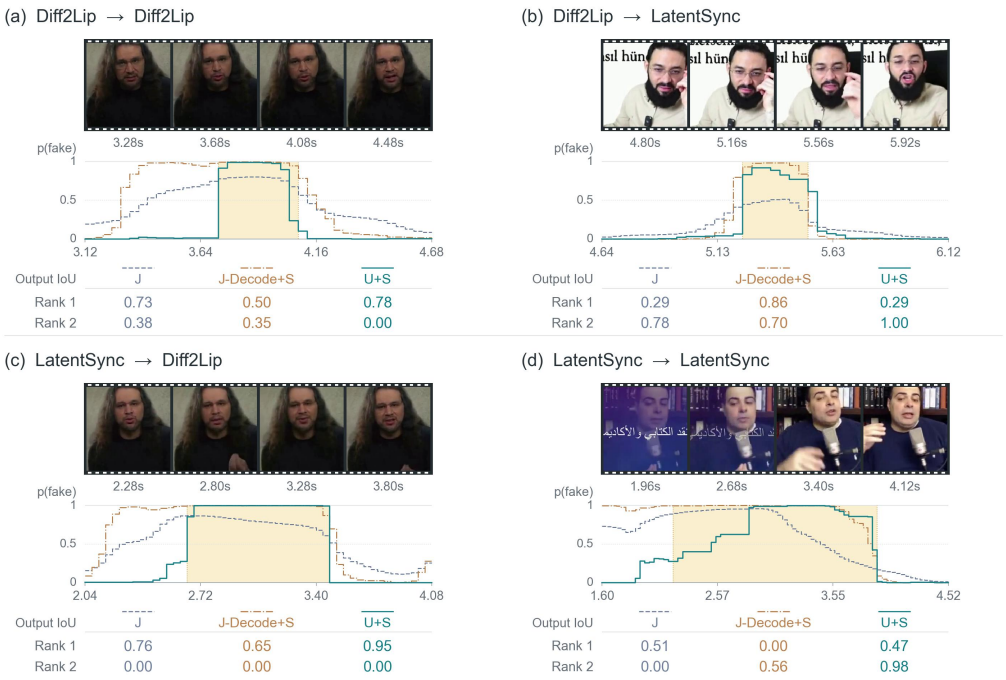


Figure 4: Fake-state probability profiles for the same four corrected directions. The frame strip, shaded ground-truth interval, three probability curves, and rank-1/rank-2 output IoUs show how structured evidence is concentrated around the target interval.

makes the effect concrete: $U + S$ produces the highest or tied-highest best-IoU interval in every corrected direction, while its probability profiles form concentrated evidence around the annotated target region.

We report the result through three linked evaluation views: the locked ArEnAV preview, corrected generator directions, and a strict visual LAV-DF transfer diagnostic. The speech factor is a designed timing prior from current-input word timestamps, with the neutral prior N providing the standard fallback. Together these views connect aggregate AP, direction-specific behavior, and interval-level evidence on the original timeline. The task is visual localization in talking-person clips, and the AP95 and compression rows characterize operating behavior at fine temporal resolution and across visual quality levels.

The matched U-max control shows how compatible-path aggregation selects a different operating point, raising segment AP50 in both generators and LatentSync AP75 while providing

a frame-level reference in Table 3. The principal system is the predeclared step-4000 $U + S$ checkpoint. External systems are included with their input and training provenance so that every comparison can be read within its corresponding evaluation view.

6 Conclusion

We presented a structured segment localizer for speech-assisted visual forgery localization. Its globally normalized training objective compares complete legal temporal labelings, while inference uses a designed current-input speech prior over the same candidate support. On the locked preview, U+S improves AP75 by 18.37 and 17.36 percentage points over source-calibrated independent segment learning for Diff2Lip- and LatentSync-trained configurations. Within this fixed scope, the result is reproducible: candidate support, decoder, scope, direction mapping, and ablations are explicit. The transfer and external-scope audits make the comparison easy to extend to complete public releases and matched author-format baselines.

7 Reproducibility, Ethics, and AI Use

All reported numbers are generated from the locked manifests and prediction files recorded with the study. README_REPRO.md, scripts/validate_numbers.py, and scripts/build_tables.py replay the locked tables; training details, checkpoint selection, candidate construction, dynamic-program recurrences, and full metric tables are provided in the appendix and scripts. Raw media, weights, and credentials remain governed by dataset and access terms.

The work analyzes publicly documented audio-visual forgery benchmarks. Because the task can affect the interpretation of media authenticity, we state the input and scope of every evaluation and keep the claims tied to those conditions. We use no human-subject intervention and show only de-identified dataset frames in the qualitative gallery.

AI-use statement. Generative AI tools assisted with literature discovery, manuscript drafting, figure code, table code, and editorial revision. The tools did not create synthetic experimental data or alter saved predictions. The authors remain responsible for verifying the numerical claims, citations, source permissions, and final scientific content.

References

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In International Conference on Digital Image Computing: Techniques and Applications, pp. 1–10, 2022. doi: 10.1109/DICTA56598.2022.10034605.

Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization. Computer Vision and Image Understanding, 236:103818, 2023. doi: 10.1016/j.cviu.2023.103818.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In Proceedings of the 32nd ACM International Conference on Multimedia, 2024. URL https://arxiv.org/abs/2311.15308.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022. URL https://arxiv.org/abs/2106.09685.

Lingpeng Kong, Chris Dyer, and Noah A. Smith. Segmental recurrent neural networks. In International Conference on Learning Representations, 2016. URL https://arxiv.org/abs/1511.06018.

Christos Koutlis and Symeon Papadopoulos. Auvire: Audio-visual speech representation reconstruction for deepfake temporal localization. In IEEE/CVF Winter Conference on Applications of Computer Vision, 2026. URL https://arxiv.org/abs/2511.18993.

Kartik Kuckreja, Parul Gupta, Injy Hamed, Thamar Solorio, Muhammad Haris Khan, and Abhinav Dhall. Tell me habibi, is it real or fake? In International Conference on Learning Representations, 2026. URL https://github.com/KjAeRsTuIsK/ArEnAV.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning, 2023. URL https://arxiv.org/abs/2212.04356.

Sunita Sarawagi and William W. Cohen. Semi-markov conditional random fields for information extraction. In Advances in Neural Information Processing Systems 17, 2004. URL https://proceedings.neurips.cc/paper/2004/hash/eb06b9db06012a7a4179b8f3cb5384d3-Abstract.html.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems, 2022. URL https://arxiv.org/abs/2203.12602.

Tianyi Wang, Xi Shao, Harry Cheng, Yinglong Wang, and Mohan Kankanhalli. Word-anchored temporal forgery localization, 2026. URL https://arxiv.org/abs/2603.06220. Submitted for review.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In European Conference on Computer Vision, volume 13664 of Lecture Notes in Computer Science, pp. 492–510, 2022.

A Additional Implementation Details

A.1 Data, splits, and exposure

The locked ArEnAV preview contains 212 evaluated videos and 106 parent groups. The two generator-specific directions contain 148 and 64 videos, respectively. The training histories use source-fit, modern-fit, and source-hold partitions recorded in the project manifests. The main checkpoints are selected by source-only validation and then locked for preview scoring. The preview is a locked evaluation scope with its training provenance recorded. Earlier training support includes LAV-DF material, so the pooled row is read together with the directional and transfer rows.

The strict LAV-DF visual scope contains 244 exact official test IDs: 128 real and 116 visual-modified rows, with 131 visual ground-truth intervals. The derived visual task uses the 128 real and 116 visual-modified rows; frozen ArEnAV-trained U rows use local-input Whisper and VideoMAE preprocessing, and the AuViRe row provides an external released-checkpoint reference.

A.2 Exact dynamic programming

For each node t, the forward recurrence sums incoming edges (s, t, z):

$$\alpha_t = \log \sum_{s,z} \exp\{\alpha_s + \log q(s, t, z) + \Psi_\theta(X, s, t, z)\},$$

with the corresponding constrained recurrence dropping edges that cross a label transition. The backward recurrence yields edge-state marginals and fake-state frame occupancy. All-real paths are retained. The speech prior changes $\log q$ but not the graph. Every unit edge has finite prior, so every native binary labeling is representable; the length cap limits edge length, not final merged output length.

The final readout thresholds fake-state frame marginals at five fixed values, forms connected runs, computes each run's mean probability, deduplicates coordinates, and applies the fixed SoftNMS setting. The readout uses the locked thresholds and checkpoint selected before preview scoring.

A.3 Training and compute

The frontend uses frozen VideoMAE blocks 0–7, rank-4 Q/V LoRA in blocks 8–11, and a shared segment head. The main selected step is 4000. The matched U-max control uses 1000 updates, seed 20260923, validation points 100/300/600/1000, and common speech readout. The locked driver records 12 physical forward computations for 26 readout conditions; the reported 261.3 seconds is a wall-clock audit on the available machine rather than a general throughput measure.

A.4 Full direction and ablation tables

Table 5: Corrected train-to-evaluation direction rows on the locked preview.

Train → eval	n	AP50	AP75	AP95	frame AP
D → D	148	43.18	30.22	1.71	55.17
D → L	64	54.61	40.65	2.32	58.07
L → D	148	34.51	24.71	1.68	56.28
L → L	64	30.73	19.50	3.33	64.32

Table 6: Matched U versus U-max control; all rows use the common speech readout.

Generator	system	AP50	AP75	AP95	frame AP	AR50
D	U frozen	24.54	14.07	2.00	45.50	33.96
D	U-max	27.05	14.14	2.10	36.62	39.34
L	U frozen	25.51	13.63	0.89	51.05	43.68
L	U-max	29.39	22.06	3.83	39.12	36.32

A.5 Robustness and external-scope details

Table 7: Corrected U+S behavior under input degradation.

Generator	condition	AP50	AP75	AP95	frame AP
D	clean	46.59	33.76	2.65	56.22
D	SNR10	47.20	36.27	4.63	58.56
D	SNR20	46.17	33.99	2.48	57.23
D	CRF23	19.27	10.82	1.88	27.86
D	CRF35	2.61	1.40	0.32	7.12
L	clean	34.03	23.91	2.62	59.34
L	SNR10	38.24	27.70	2.49	60.52
L	SNR20	35.62	25.38	2.31	59.83
L	CRF23	6.46	4.12	0.56	13.35
L	CRF35	0.43	0.03	0.00	5.33

A.6 Qualitative selection rule

The visual audit uses the same four corrected train-to-evaluation directions throughout Figures 3 and 4. Figure 3 reports the sampled frames, ground-truth interval, three interval predictions, and each system's best IoU. Figure 4 reports the corresponding fake-state probability profiles together with rank-1 and rank-2 output IoUs. Video frames are extracted from the saved media at the displayed timestamps; interval coordinates and IoU values come directly from the locked prediction files. The displayed frames come from the saved media at the stated timestamps, while interval coordinates and IoU values come directly from the locked prediction files.