AV-FABLE: AUDIO-VISUAL MANIPULATION FORENSICS IN LONG-FORM VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Localized AI manipulations in long-form videos often affect brief events within otherwise authentic content, making video-level classification insufficient for forensic analysis. Existing audio-visual deepfake benchmarks largely focus on short talking-head clips, while long-form temporal forensics remains predominantly visual. We introduce AV-FABLE, a dual-timeline benchmark for sparse manipulations in open-scene long-form audio-visual videos. AV-FABLE contains 16,650 pristine–manipulated twin pairs (998 hours) created with 27 editing configurations, with separate audio and visual timelines that enable modality-specific temporal localization and expose errors hidden by joint-timeline evaluation. On AV-FABLE, zero-shot prompting of audio-visual large language models (AV-LLMs) yields near-chance detection and poor temporal localization. In contrast, supervised readouts of their frozen perception features recover strong forensic signals, suggesting manipulation evidence is encoded in these representations but not effectively exposed through language outputs. Motivated by this observation, we propose NFSR (Native Forensics via Separated Readouts), which reads frozen AV-LLM features with separate audio and visual temporal decoders. NFSR achieves the best audio-visual localization performance in-domain and across all three out-of-distribution settings, including transfer to held-out videos from an external long-video forensic benchmark. AV-FABLE, code, and model weights will be publicly released.

1 INTRODUCTION

A few manipulated seconds can change the meaning of an otherwise authentic video. Replacing an utterance can alter what a speaker appears to have said; inserting a sound can suggest an event that never occurred; modifying a brief action can change the interpretation of an entire sequence. Such edits may affect the audio, the visual stream, or both, often with different temporal boundaries. A video-level authenticity label therefore reveals little about where a manipulation occurs or which modality is affected. Fine-grained forensic analysis requires identifying **when a manipulation occurred and which modality was affected**.

Existing benchmarks capture parts of this problem, but not this setting as a whole. LAV-DF and the AV-Deepfake1M series advance audio-visual deepfake detection and temporal localization in predominantly speaker-centered videos (Cai et al., 2022; 2024; 2025). TASLE and ActivityForensics extend visual forensics to localized manipulations in broader, untrimmed videos (Feng et al., 2026; Bao et al., 2026), while AVID studies audio-visual inconsistency understanding and temporal grounding in long-form videos (Chen et al., 2026b). What remains missing is a benchmark that localizes manipulations independently on the audio and visual timelines of open-scene long-form video. This distinction is important because audio and visual edits may overlap, occur asynchronously, or affect only one modality. We target this setting: sparse manipulations of speech, environmental sounds, actions, objects, and attributes embedded within largely authentic long-form content.

The distinction between modalities is central to this setting. Audio and visual manipulations may overlap in time, occur asynchronously, or affect only one modality. A joint timeline can therefore reward a prediction for finding the right moment even when it identifies the wrong stream. Cross-modal consistency is also insufficient: two streams can remain plausible even when one has been altered. Evaluation should therefore preserve both temporal boundaries and modality identity.

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.

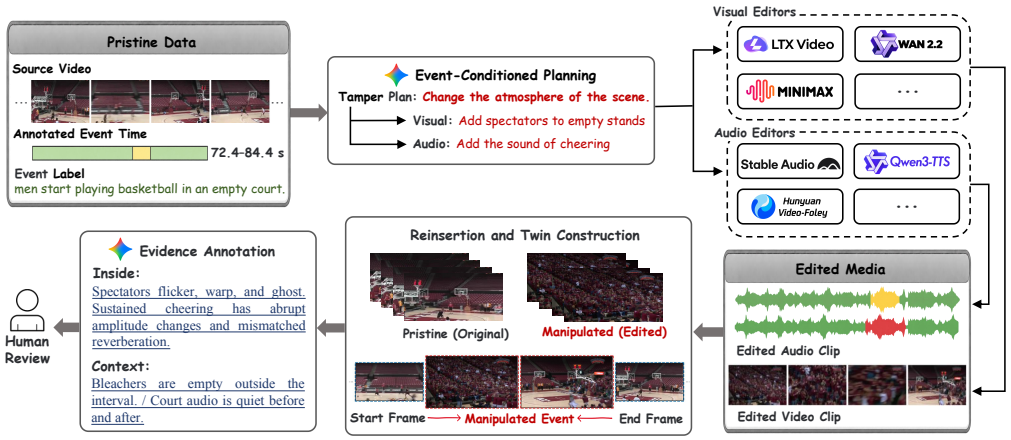


Figure 1: **Overview of SEE-Forge, the event-driven pipeline for constructing AV-FABLE.** Given an event in a pristine video, SEE-Forge plans audio and/or visual manipulations, executes them with editors, and reinserts them into context to create pristine–manipulated twin pairs. The executed interval provides temporal ground truth, while human-reviewed annotations record forensic cues within and around the manipulated event.

We introduce **AV-FABLE** (Forensic Audio-visual Benchmark for Long-form Events), a dual-timeline benchmark for sparse manipulations in open-scene long-form audio-visual videos. AV-FABLE comprises **16,650 pristine–manipulated twin pairs spanning 998 hours and 27 editing configurations**; videos are minute-scale (median 88 s), and manipulations cover a median of 9.45% of each manipulated video. Our construction pipeline, **SEE-Forge**, selects natural events in the source footage, plans a local audio or visual edit for each, routes the edit to a suitable generator, and reinserts the result into the original video (Figure 1). Each manipulated video has a pristine twin from the same source under a matched export process, reducing content and processing confounds. Separate audio and visual timelines record the interval and modality of every executed edit, and human-verified evidence records describe how each edit can be observed. The benchmark centers on temporal localization, together with detection, modality attribution, and explanation, and evaluates transfer to held-out sources, generators, and editing routes with thresholds fixed on source-domain data. Native-track evaluation exposes failures that joint-timeline evaluation hides: the visual predictions of our model alone reach 64.04 Joint-mAP, above every visual-only baseline, while missing every audio event.

When prompted zero-shot on AV-FABLE, the audio-visual large language models (AV-LLMs) Qwen2.5-Omni and Qwen3-Omni (Xu et al., 2025a;b) reach only 51–52 detection AUROC and at most 0.30 mAP on either track. Yet linear probes on their frozen perception features separate manipulated from authentic segments with 90.8 (audio) and 80.8 (visual) AUROC. This striking gap suggests that AV-LLMs encode useful forensic evidence in their perceptual representations, but fail to reliably expose it through zero-shot language outputs. This observation motivates learning temporal readouts directly from these frozen features.

Building on this finding, we propose **NFSR** (Native Forensics via Separated Readouts), which directly reads frozen audio and visual representations for modality-specific temporal localization. NFSR combines intermediate-layer and interface features, conditions local evidence on the host video, and uses separate temporal decoders to localize audio and visual manipulations. Their predictions further support video-level detection and modality attribution. In-domain, NFSR achieves 91.30 AV-mAP and 96.04 detection AUROC, compared with 84.76 AV-mAP for UMMAFormer using the same frozen features. NFSR further achieves the highest AV-mAP across all three OOD settings, including transfer to a source-disjoint subset of the external TASLE benchmark. Our contributions are summarized as follows:

- **AV-FABLE and the dual-timeline task:** a large-scale benchmark for sparse audio-visual temporal localization in long-form videos, with pristine–manipulated twin pairs, separate audio and visual annotations, and three OOD settings. Its native-track evaluation exposes failures hidden by joint-timeline evaluation.

- **A finding on AV-LLMs:** zero-shot prompting remains near chance on AV-FABLE, while supervised readouts of frozen perception features recover strong forensic signals, revealing a gap between internal representations and language outputs.
- **NFSR:** a readout of frozen AV-LLM encoders that combines intermediate-layer and interface features, host conditioning, and separate audio and visual decoders. NFSR achieves the highest AV-mAP in-domain and across all three OOD settings.

2 Related Work

Benchmarks for manipulation detection and temporal localization. Facial-forgery benchmarks primarily assess image- or video-level authenticity (Rössler et al., 2019; Li et al., 2020; Dolhansky et al., 2020), with FakeAVCeleb adding manipulated speech to talking-head video (Khalid et al., 2021). ForgeryNet and PartialSpoof provide interval-level supervision for visual and audio manipulations, respectively (He et al., 2021; Zhang et al., 2023a). LAV-DF and the AV-Deepfake1M series extend temporal localization to content-driven audio-visual edits, predominantly in speaker-centered clips (Cai et al., 2022; 2023; 2024; 2025). Beyond this setting, AVFakeBench broadens scene coverage (Xia et al., 2026), while Omni-Fake brings several media types into a common forensic benchmark (Li et al., 2026b). Longer authentic context is central to TASLE and ActivityForensics, which embed sparse visual manipulations within untrimmed or activity-rich videos (Feng et al., 2026; Bao et al., 2026). AVID studies long-form audio-visual inconsistency grounding (Chen et al., 2026b), but cross-modal compatibility and the authenticity of each stream remain distinct targets. AV-FABLE combines open-scene long-form videos with separate temporal annotations and native-track evaluation, requiring both manipulation boundaries and the affected modality to be identified.

Temporal localization and forensic representations. Beyond classification (Chollet, 2017; Jung et al., 2022; Chen et al., 2026a), temporal forensics studies boundary prediction in audio (Wu et al., 2024; Zhong et al., 2024) and audio-visual content (Cai et al., 2023; Chen et al., 2025b; Shen et al., 2026). Audio-visual localizers exploit cross-modal cues (Koutlis & Papadopoulos, 2026; Anshul et al., 2025), adaptive features in UMMAFormer (Zhang et al., 2023b), or query-based and contextual modeling (Zhu et al., 2025; Yin et al., 2026). Action detection supplies interval predictors (Zhang et al., 2022), while TASLE's MSLoc uses language-assisted proposal refinement (Feng et al., 2026). Modality-decoupled detectors retain stream-specific evidence (Peng et al., 2026; Li et al., 2026a). AV-LLMs provide language-based predictions (Xu et al., 2025a;b; Tang et al., 2025), whereas supervised probes assess forensic information accessible from pretrained features (Alain & Bengio, 2016; Boldisor et al., 2026). NFSR conditions frozen perception features on the input host video and learns separate audio and visual temporal readouts. Appendix B expands these comparisons and reviews generation and editing techniques.

3 AV-FABLE

Table 1 compares AV-FABLE with representative forensic benchmarks: among them, only AV-FABLE combines open-scene long-form videos, separate audio and visual timelines, pristine twins, and event-grounded explanations. It comprises **16,650 twin pairs** (33,300 videos; 998.1 hours), with 10,306 audio and 11,847 visual events across 27 editing configurations.

3.1 Benchmark Design and Task

For a video $X = (A, V)$ of duration T, each modality $m \in \{a, v\}$ carries its own event set $\mathcal{G}_m = \{(b_i^m, e_i^m, o_i^m, k_i^m)\}_{i=1}^{N_m}$, where b_i^m, e_i^m are temporal boundaries, o_i^m is the manipulation operation, and k_i^m is the target category. Audio and visual events may overlap, occur asynchronously, or appear on only one track. Given X, a system returns confidence-scored intervals on each track (localization), a video-level authenticity score (detection), one of four attribution classes—clean, audio-only, visual-only, or both-track—and an evidence statement for each predicted event (explanation).

Native-track evaluation preserves modality identity: a visual prediction cannot match an audio event, even at identical timestamps. It is the primary localization protocol of AV-FABLE, summarized by AV-mAP, the mean of A-mAP and V-mAP. *Joint-timeline evaluation* instead merges overlapping ground-truth intervals across tracks and matches predictions regardless of their track (Joint-mAP);

Table 1: **Forensic benchmarks comparison.** Length: short clips (Clip) or minute-scale videos (Long). Manip.: audio (A) or visual (V) manipulations; Loc.: temporal localization; Expl.: explanations; Pair: pristine–manipulated twins; OOD shift: held-out sources, generators, or manipulation types/routes. (✓): partial support, or shift without a dedicated held-out split.

Dataset	Scale	Scene	Length	Manip.		Loc.	Expl.	Pair	OOD shift		
				A	V				Src	Gen	Type
FaceForensics++ (Rössler et al., 2019)	6,000 v	Faces	Clip	×	✓	×	×	(✓)	×	(✓)	(✓)
Celeb-DF (v2) (Li et al., 2020)	6,529 v	Celebrity faces	Clip	×	✓	×	×	(✓)	×	×	×
DFDC (Dolhansky et al., 2020)	128,154 v	Consented actors	Clip	(✓)	✓	×	×	(✓)	(✓)	(✓)	×
FakeAVCeleb (Khalid et al., 2021)	20,000 v	Talking heads	Clip	✓	✓	×	×	(✓)	×	×	×
ForgeryNet (He et al., 2021)	221,247 v	Faces	Clip	×	✓	✓	×	(✓)	×	×	×
LAV-DF (Cai et al., 2022; 2023)	136,304 v	Talking heads	Clip	✓	✓	✓	×	(✓)	×	×	×
AV-Deepfake1M++ (Cai et al., 2025)	2,051,154 v	Talking heads	Clip	✓	✓	✓	×	(✓)	(✓)	(✓)	×
TASLE (Feng et al., 2026)	12,472 v	Open scenes	Long	×	✓	✓	✓	(✓)	(✓)	✓	×
ActivityForensics (Bao et al., 2026)	5,389 v	Open activities	Long	×	✓	✓	×	(✓)	×	✓	(✓)
AV-FABLE (ours)	**33,300 v**	**Open scenes**	**Long**	✓	✓	✓	✓	✓	✓	✓	✓

we report it as an auxiliary view. Comparing the two separates locating a manipulated interval from identifying the track it affects. Predictions on pristine videos count as false positives under both protocols, as do predictions on an unmodified track under native-track evaluation.

3.2 Event-Driven Construction and Annotation

From natural events to executable edits. SEE-Forge (Figure 1) edits events that already occur in the host video, i.e., the source video that receives the edit, rather than arbitrary time windows. ActivityNet (Heilbron et al., 2015) and EPIC-KITCHENS-100 (Damen et al., 2022) provide third-person activities and egocentric interactions; HDTF (Zhang et al., 2021) supplies held-out talking-head videos. For each host video, we build an event catalog from captions, speech transcripts, acoustic onsets, motion, and object interactions. Gemini 3.7 Flash (Google DeepMind, 2026) then acts as the planner: it selects one cataloged event and specifies the target track, interval, semantic change, and content to preserve. A second model verifies that the target is observable and that a compatible generator can execute the instruction.

Editing and pristine twins. Each of the 27 editing configurations pairs a generator with an execution route: frame-conditioned generation, masked or unmasked video-to-video editing, joint audio-visual generation, audio generation, or speech replacement (Table 6). SAM 2 masks constrain localized visual edits (Ravi et al., 2025). Generators receive the target interval with 2–5 s of surrounding context, and their outputs are reinserted at the recorded coordinates with checks on motion continuity, visual anchors, and loudness at the seams. Each manipulated video has a pristine twin from the same host video, and both twins pass through the same cropping, reinsertion, and export steps (resolution, frame rate, codec, and audio format). These controls address the content and processing shortcuts reported for earlier datasets (Smeu et al., 2025): metadata-only classifiers reach at most 54.6 AUROC (Table 8). Traces that generators leave inside the edited interval, such as local resampling, generator texture, or synthesis residue, are beyond the reach of file-level probes and are discussed in Appendix A.5. The pristine twin also serves as a content-matched reference for annotation and error analysis.

Temporal labels and observable evidence. Temporal labels are execution-grounded: each event records the track, interval, operation, and target category of an executed edit, and "ground truth" refers to these annotations. Intervals mark where edits were applied rather than perceptual boundaries, so a perceptible change may cover only part of an interval or begin and end slightly off its boundaries; reviewers check boundaries during retention but do not re-draw them. Operations cover replacement, insertion, regeneration, and removal; targets include speech, sounds, actions, objects, and scene content. For 20,529 events, Gemini 3.7 Flash drafts a candidate evidence record from a short clip around the event. Evidence records distinguish observations *inside* the event, *context* from surrounding or paired media, and the *judgment* these observations support, so the intended edit and its observable trace are recorded separately. Automatic checks screen decoding, synchronization, and integration failures. Two reviewers then independently inspect all 22,153 events against the pristine twin and the edit instruction. They decide whether the edit was realized within its recorded interval without collateral changes, abrupt boundary transitions, or visual breakdown (Cohen's $\kappa = 0.84$) (Cohen, 1960) and whether the media support its evidence statements ($\kappa = 0.82$); a

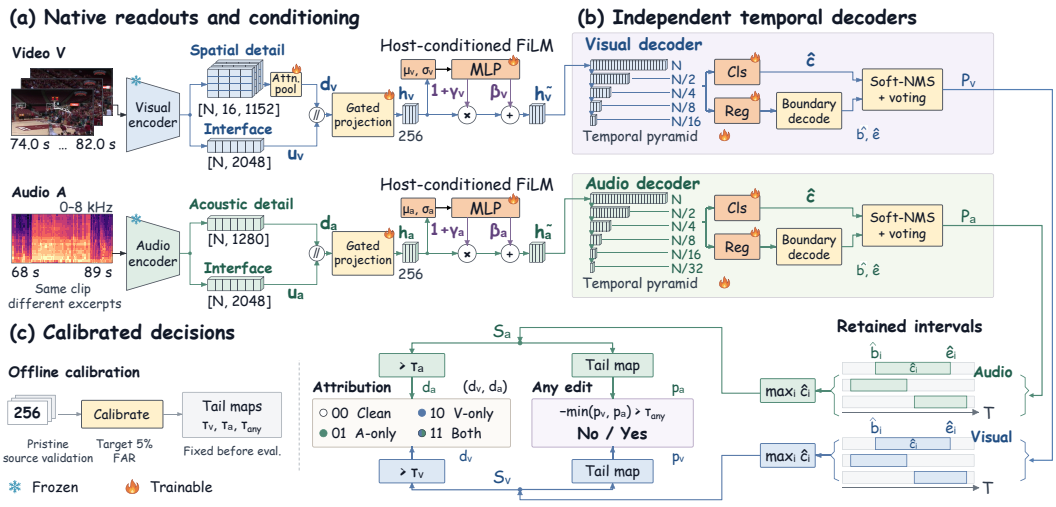


Figure 2: **Overview of NFSR.** (a) Frozen intermediate-layer ("detail") and interface features pass through modality-specific gated projection and host-conditioned FiLM. (b) Independent decoders localize audio and visual manipulations on separate timelines. (c) Source-calibrated thresholds give video-level detection and four-way attribution without changing the intervals. Audio and video excerpts come from the same clip; intervals are schematic.

third reviewer adjudicates the 7.7% of events with disagreements (Appendix A.4). Events without supported evidence remain available for detection and localization but not explanation.

3.3 Dataset Characteristics and Evaluation Splits

The dataset contains 4,903 audio-only, 6,376 visual-only, and 5,371 both-track manipulated videos. Across manipulated videos, the union of audio and visual event intervals covers a median of 9.45% of the video, while 91.5% have at most 30% manipulated time. Videos are minute-scale: the median lasts 88 s and 80% last between 33 and 207 s. The median event lasts about 9 s in the training and ID splits, so each manipulation occupies a small part of its timeline. In 561 both-track videos, the audio and visual events do not overlap, so each track requires its own boundaries.

We allocate 10,954/2,347/2,347/1,002 twin pairs to training, validation, in-domain (ID) testing, and OOD testing; all derivatives of a host video remain in the same split. Of the 27 editing configurations, 23 appear in training and four are held out. We evaluate three OOD settings (Section 5.2): *source transfer* to HDTF and a source-disjoint TASLE subset, *held-out generators* on seen routes, and *held-out TI2V*, a text-and-image-conditioned video generation route withheld together with its generator. Evaluation covers localization, detection, attribution, and explanation, with decision thresholds fixed on source-validation data (Section 5). Explanation is the most challenging dimension: current baselines produce well-formed outputs but often misidentify the operation or target (Appendix H), leaving substantial room for future work. Appendices A and C detail the composition, construction settings, and split protocols.

4 NFSR: Native Forensics via Separated Readouts

NFSR reads frozen perception features and localizes manipulations on separate audio and visual timelines. Each of its three components addresses one observation: supervised readouts recover discriminative signals from perception features below the language interface (Section 4.1) whether a local change is anomalous depends on its host video (Section 4.2); and the two modalities need not share manipulation boundaries (Section 4.3). Figure 2 summarizes the framework.

4.1 Reading Frozen Perception Features

Zero-shot Qwen-Omni models remain near chance for detection through the language interface, whereas linear probes on their frozen perception features separate manipulated from authentic

segments (Tables 2 and 15). NFSR therefore reads the perception features directly. We instantiate it with the frozen audio and visual encoders of Qwen3-Omni (Xu et al., 2025b). For each modality $m \in \{a, v\}$, we extract *intermediate-layer features* from a fixed intermediate layer and *interface features* from the encoder output passed to the language model. The former retain fine perceptual detail, while the latter provide a complementary representation aligned with the language interface.

For a video of duration T, both feature types are aligned to $N = \lceil T/\Delta \rceil$ positions with $\Delta = 0.25\,\mathrm{s}$. Visual spatial tokens are attention-pooled at each position; audio features are temporal vectors. Denoting the intermediate-layer and interface features by $\mathbf{d}_t^m$ and $\mathbf{u}_t^m$, a modality-specific gated projection forms $\mathbf{h}_t^m = \phi_m(\mathbf{d}_t^m, \mathbf{u}_t^m) \in \mathbb{R}^D$, with $D = 256$. Reading both layers lets the temporal model combine their information rather than committing to a single representation depth.

4.2 Host-Conditioned Feature Modulation

A local sound, motion, or texture change takes on different significance across video conditions and scene content. With sparse manipulation, the extended host video provides context for interpreting such changes. We therefore condition local features on the input video's channel-wise mean $\boldsymbol{\mu}_m$ and standard deviation $\boldsymbol{\sigma}_m$, computed over valid temporal positions. Following feature-wise linear modulation (Perez et al., 2018), we apply

$$(\boldsymbol{\gamma}_m, \boldsymbol{\beta}_m) = g_m([\boldsymbol{\mu}_m; \boldsymbol{\sigma}_m]), \qquad \widetilde{\mathbf{h}}_t^m = \mathbf{h}_t^m \odot (\mathbf{1} + \boldsymbol{\gamma}_m) + \boldsymbol{\beta}_m. \tag{1}$$

Here, g_m is a two-layer MLP, $[\cdot;\cdot]$ denotes concatenation, and $\odot$ is element-wise multiplication. The scale and shift vectors $\boldsymbol{\gamma}_m, \boldsymbol{\beta}_m \in \mathbb{R}^D$ are shared across positions within a video but vary with its context. The MLP thus learns how host statistics should reweight local evidence while keeping the features' absolute scale, which per-video standardization would discard. The statistics come from the input video alone, in training and at inference.

4.3 Modality-Separated Temporal Localization

An event in one modality does not establish manipulation in the other. We use independently parameterized decoders D_a and D_v, each receiving only its own conditioned feature sequence. This keeps the evidence path modality-specific while allowing different event boundaries on the two tracks. Following multi-scale temporal localization (Zhang et al., 2022; Yang et al., 2024), each decoder constructs a temporal pyramid with masked dilated convolutions and predicts a classification logit and nonnegative left/right distances at each point:

$$(\hat{b}_j^m, \hat{e}_j^m, \hat{c}_j^m) = \left(u_j - r_j \hat{d}_j^{m,-},\ u_j + r_j \hat{d}_j^{m,+},\ \mathrm{sigmoid}(z_j^m)\right). \tag{2}$$

Here, j indexes points across pyramid levels, u_j is the point's timestamp, and r_j is its level's temporal stride in seconds. The decoder outputs z_j^m and distances $\hat{d}_j^{m,-}, \hat{d}_j^{m,+}$ in stride units. Thus, the first two terms recover event boundaries in video time, and the third gives classification confidence. Temporal Soft-NMS (Bodla et al., 2017) combines proposals across scales; surviving intervals are clipped to video bounds to form $\widehat{\mathcal{S}}_m$.

Training objective. Each branch is supervised by its own temporal annotations:

$$\mathcal{L}_{\mathrm{readout}} = \sum_{m \in \{a,v\}} (\mathcal{L}_{\mathrm{focal}}^m + \lambda_{\mathrm{reg}} \mathcal{L}_{\mathrm{DIoU}}^m). \tag{3}$$

Here, $\mathcal{L}_{\mathrm{focal}}^m$ is sigmoid focal loss (Lin et al., 2017) over valid temporal points, and $\mathcal{L}_{\mathrm{DIoU}}^m$ is one-dimensional distance-IoU loss (Zheng et al., 2020) on points assigned to ground-truth events; $\lambda_{\mathrm{reg}} = 1$ balances the terms. Pristine videos and unmodified tracks supply negative supervision. Assignment rules and optimization settings appear in Appendix C.

Detection and attribution. Each track's video score S_m is its highest proposal confidence. We convert S_m to an upper-tail probability p_m relative to 256 pristine source-validation videos and score any-manipulation detection as $S_{\mathrm{any}} = -\min\{p_a, p_v\}$. Detection and each track decision use their own threshold, calibrated once at a 5% false-alarm budget and reused on every test split. The two track decisions give the attribution class (clean, audio-only, visual-only, or both-track) without altering either event set; Appendix C gives the full scoring rules.

Table 2: **ID main results on AV-FABLE (%).** Blocks are visual-only, audio-only, and audio-visual. Outputs are separate audio/visual intervals (A/V), one track (A or V), fixed-window intervals (win), modality-labeled language timestamps (Text), or modality-unlabeled intervals (Unified). – marks unsupported metrics; absent tracks count as misses in AV-mAP and Joint-timeline metrics. NFSR, UMMAFormer, and TASLE-trained ActionFormer are three-seed means (Table 18); others are single runs. **Bold** and underlined values mark the best and second-best results per column; FAR is an operating-point diagnostic and is not ranked.

Method	Training Form	Output	AUROC	FAR	Attr. F1	AV-mAP	A-mAP	V-mAP	Joint-mAP	A-IoU	V-IoU	Joint-IoU
Xception	Official weights	V_{win}	50.71	4.94	17.62	0.01	–	0.02	0.02	–	0.22	0.24
MSLoc-PG	Official weights	V	57.34	3.54	20.37	0.95	–	1.90	1.36	–	4.55	3.40
ActionFormer	TASLE	V	67.94	2.78	20.99	5.06	–	10.13	7.56	–	5.56	4.55
Xception	AV-FABLE	V_{win}	66.27	4.94	23.27	0.04	–	0.08	0.06	–	4.87	3.98
DeMamba (Stage-1 window)	AV-FABLE	V_{win}	74.35	4.26	28.77	24.75	–	49.49	26.83	–	35.21	28.54
ActionFormer	AV-FABLE	V	84.55	5.84	34.95	44.61	–	89.22	59.93	–	**80.66**	62.71
AASIST	Official weights	A_{win}	55.17	5.07	18.65	0.00	0.01	–	0.04	0.51	–	0.38
AASIST	AV-FABLE	A_{win}	75.95	5.62	29.42	0.03	0.06	–	0.07	16.26	–	12.06
Qwen2.5-Omni-7B (ZS)	Official weights	Text	51.04	0.04	10.92	0.28	0.25	0.30	0.26	0.00	0.00	0.00
Qwen3-Omni-30B-A3B (ZS)	Official weights	Text	52.28	4.13	20.42	0.03	0.02	0.04	0.03	0.56	0.47	0.60
video-SALMONN 2+ 7B (ZS)	Official weights	Text	49.98	0.43	0.59	0.00	0.00	0.00	0.00	0.01	0.02	0.02
DAV-Det	Official weights	A/V_{win}	56.52	4.90	22.30	0.40	0.59	0.21	0.57	2.90	0.86	2.68
BA-TFD+	AV-FABLE	A/V	66.71	4.52	27.07	0.88	1.54	0.22	0.63	37.54	1.23	23.58
UMMAFormer + Qwen3	AV-FABLE	A/V	89.18	5.75	74.83	84.76	89.90	79.62	58.42	79.77	65.25	67.44
Qwen3 direct timestamp SFT	AV-FABLE	Text	76.53	4.35	47.66	37.80	57.30	18.30	19.20	56.06	26.86	33.99
AuViRe	AV-FABLE	Unified	64.87	6.31	–	–	–	–	0.42	–	–	11.21
ICS-Sync	AV-FABLE	Unified	59.61	6.43	–	–	–	–	5.47	–	–	8.39
DAV-Det + AV-FABLE	AV-FABLE	A/V_{win}	80.56	4.05	48.02	1.50	0.80	2.21	1.51	10.49	16.85	14.79
NFSR + Qwen3 (Ours)	AV-FABLE	A/V	**96.04**	3.52	**83.61**	**91.30**	**92.24**	**90.36**	**72.33**	**84.72**	76.90	**81.21**

5 EXPERIMENTS

We evaluate whether frozen perception features support sparse audio-visual temporal localization, how the resulting readouts transfer to the three OOD settings, and how feature depth, host conditioning, and modality separation affect these results.

Evaluation protocol. We use AV-FABLE's source-disjoint splits. Localization uses confidence-ranked mAP@[0.30:0.05:0.70] and Event-IoU, the mean best overlap per event, with zero assigned to misses. AV-mAP, headline measure, averages A-mAP and V-mAP over tracks with events; a method without output on a track receives zero AP there. Joint-timeline metrics (Joint-mAP and Joint-IoU; Section 3.1) apply to outputs without track labels, but count a correct co-timed prediction on the second track as a duplicate, so we report them as an auxiliary view (Appendix C.3). Detection uses AUROC and realized false-alarm rate (FAR). Attribution reports macro-F1 over the four classes (clean, audio-only, visual-only, both-track) present in each evaluation set's ground truth (Eq. 5); held-out TI2V, for example, contains only clean and visual-only videos, so attribution is compared within rather than across settings. Thresholds are set on 256 pristine source-validation videos at a nominal 5% FAR budget and held fixed on ID and OOD data (Appendix C).

Baselines and implementation. We compare three prediction approaches: language-interface inference with Qwen2.5-Omni-7B, Qwen3-Omni-30B-A3B, and video-SALMONN 2+ 7B (Xu et al., 2025a;b; Tang et al., 2025); single-modality detection with Xception (Chollet, 2017), AASIST (Jung et al., 2022), and DeMamba (Chen et al., 2026a); and temporal or audio-visual modeling with BA-TFD+ (Cai et al., 2023), ActionFormer (Zhang et al., 2022), UMMAFormer (Zhang et al., 2023b), AuViRe (Koutlis & Papadopoulos, 2026), ICS-Sync (Anshul et al., 2025), and DAV-Det (Peng et al., 2026). Qwen3-Omni timestamp SFT additionally tests supervised language-based localization. We include the released MSLoc proposal generator (Feng et al., 2026); its full pipeline and baseline adaptation details appear in Appendix F and Appendix C, respectively. These groups differ in role and output form (Table 2): AV-LLMs test whether prompted outputs expose manipulations, single-modality detectors are track-specific references, and localizers with separate audio and visual interval outputs are the direct comparison. For fixed-window, single-track, unified, and text outputs, localization scores also reflect the fit of the output form to the task; AV-FABLE-trained AASIST, for example, reaches 75.95 AUROC but 0.06 A-mAP after window-to-interval conversion. NFSR trains 256-dimensional readouts on frozen Qwen3-Omni features at 0.25 s resolution for 12 epochs using

Table 3: **OOD performance (%).** Metrics, source-calibrated thresholds, rows, –, and bold/underline follow Table 2. For source transfer, detection averages HDTF/TASLE, attribution and V-mAP pool them, and A-mAP uses HDTF (384 audio events). Held-out TI2V has visual edits only (AV-mAP = V-mAP). Attribution F1 uses three, four, and two classes, respectively. †TASLE overlaps the model's training distribution. Joint-timeline metrics and Event-IoU: Table 12.

Method	Training Form	Source transfer						Held-out generators						Held-out TI2V			
		AUROC	FAR	Attr. F1	AV-mAP	A-mAP	V-mAP	AUROC	FAR	Attr. F1	AV-mAP	A-mAP	V-mAP	AUROC	FAR	Attr. F1	AV-mAP
Xception	Official weights	48.41	3.53	22.27	0.30	–	0.59	50.59	5.47	18.96	0.11	–	0.23	50.18	5.33	37.21	0.25
MSLoc-PG	Official weights	72.37	0.97	47.97	26.51	–	53.01†	57.86	3.22	22.82	3.82	–	7.65	56.88	5.00	40.18	0.92
ActionFormer	TASLE	73.82	0.00	47.55	37.50	–	**75.00**†	77.50	2.10	24.05	10.52	–	21.05	66.78	2.89	44.00	15.15
Xception	AV-FABLE	52.63	1.60	24.24	0.15	–	0.31	53.26	3.86	19.44	0.25	–	0.49	53.91	4.33	38.46	0.67
DeMamba (Stage-1 window)	AV-FABLE	54.94	2.23	25.65	0.12	–	0.23	55.75	1.93	18.16	1.18	–	2.37	59.66	4.33	39.11	2.17
ActionFormer	AV-FABLE	62.31	3.74	40.01	19.13	–	38.26	69.10	4.50	34.26	25.60	–	**51.21**	**85.46**	6.67	**72.03**	61.57
AASIST	Official weights	52.43	0.00	22.24	0.00	0.00	–	52.08	5.14	19.20	0.04	0.07	–	50.00	4.67	32.80	0.00
AASIST	AV-FABLE	71.61	1.04	52.46	0.03	0.05	–	61.70	3.22	26.29	0.27	0.55	–	50.00	6.67	32.56	0.00
Qwen2.5-Omni-7B (ZS)	Official weights	48.92	0.00	10.96	0.92	0.34	1.50	49.11	0.00	10.27	0.62	0.56	0.68	51.30	0.00	18.45	0.86
Qwen3-Omni-30B-A3B (ZS)	Official weights	51.98	7.70	35.41	1.14	0.01	2.26	50.98	4.50	19.99	0.04	0.05	0.04	51.31	6.00	32.56	0.12
video-SALMONN 2+ 7B (ZS)	Official weights	50.02	0.00	22.24	0.00	0.00	0.00	50.00	0.32	18.50	0.00	0.00	0.00	49.83	0.33	33.30	0.00
DAV-Det	Official weights	64.78	24.03	45.68	10.47	0.10	20.84	53.23	4.50	20.50	0.81	0.56	1.07	51.24	3.67	38.41	1.27
BA-TFD+	AV-FABLE	52.43	2.21	23.99	0.41	0.65	0.17	57.23	3.22	21.88	0.59	1.09	0.09	50.00	6.00	34.84	0.24
UMMAFormer + Qwen3	AV-FABLE	78.81	12.21	71.78	60.48	**98.91**	22.05	68.71	4.18	35.23	33.98	40.98	26.97	63.46	5.00	48.01	36.06
Qwen3 direct timestamp SFT	AV-FABLE	76.40	0.66	47.16	23.94	47.78	0.09	59.52	6.11	31.16	10.80	21.30	0.30	51.63	4.33	34.61	0.26
AuViRe	AV-FABLE	51.99	1.04	–	–	–	–	54.82	6.75	–	–	–	–	50.36	8.67	–	–
ICS-Sync	AV-FABLE	50.56	1.30	–	–	–	–	54.10	6.75	–	–	–	–	49.43	7.00	–	–
DAV-Det + AV-FABLE	AV-FABLE	61.52	40.45	40.79	8.60	0.14	17.05	60.06	5.47	25.49	3.36	1.15	5.56	63.32	4.00	43.75	10.66
NFSR + Qwen3 (Ours)	AV-FABLE	**89.86**	2.97	**77.21**	**70.15**	98.37	41.92	**79.40**	3.86	**49.06**	**49.12**	**48.95**	49.29	84.56	4.67	68.06	**68.07**

AdamW (10^{-4} learning rate) (Loshchilov & Hutter, 2017); UMMAFormer shares its features and training budget. All architecture choices, layer choices, hyperparameters, and model selection were determined without access to OOD test results; checkpoints are selected by source-validation mAP.

5.1 Forensic Readout and In-Domain Results

Supervised readouts recover signals that prompted outputs miss. Zero-shot AV-LLM predictions yield near-chance detection and at most 0.30 mAP on either track (Table 2), partly because of their text-timestamp output. Timestamp SFT keeps this output form and raises A-mAP to 57.30 but V-mAP only to 18.30. On the same frozen features, linear probes outperform unsupervised feature-deviation and temporal-jump scores on both tracks (Table 15), and NFSR reaches 92.24 A-mAP and 90.36 V-mAP. The forensic evidence is therefore accessible to supervised readouts even when prompted outputs do not reveal it.

NFSR improves localization on shared frozen features. With the same Qwen3 features and training budget, NFSR raises AV-mAP over UMMAFormer from 84.76 to 91.30, the highest in Table 2, with the largest gain on the visual track (+10.74 V-mAP). It also improves detection AUROC by 6.86 points and Event-IoU on both tracks, so the gain covers event coverage as well as ranking. NFSR also achieves the highest attribution F1.

Native-track evaluation exposes missing modalities. The AV-FABLE-trained ActionFormer is competitive on the visual track (89.22 V-mAP and higher V-IoU than NFSR); its lower AV-mAP reflects its visual-only output rather than weaker visual localization. Conversely, NFSR's visual predictions alone reach 64.04 Joint-mAP, above every visual-only baseline, despite zero A-mAP (Table 11); only native-track evaluation exposes this missing modality.

5.2 Transfer Across OOD Settings

We evaluate the three OOD settings of Section 3.3 (composition in Table 10) without target-domain adaptation or threshold adjustment. Source transfer changes both domain and modality composition, and held-out TI2V changes both route and generator; each result therefore measures transfer to the setting as a whole.

Localization gains hold in OOD settings. NFSR achieves the highest AV-mAP in every setting (Table 3) and exceeds the matched-feature UMMAFormer, mainly through higher V-mAP; the two are on par on HDTF audio. NFSR also has higher attribution F1 and Event-IoU than UMMAFormer in all three settings (Table 12). Visual-only baselines remain competitive on their own track: the AV-FABLE-trained ActionFormer slightly exceeds NFSR in V-mAP on held-out generators and

Table 4: **NFSR core component ablation.** Three-seed mean ± sample standard deviation with matched 12-epoch training, splits, scoring, and thresholds. OOD-All averages Table 3; A-mAP uses its two audio-edit settings (Appendix C). Bold/underline mark the best and second-best means (FAR not ranked). Joint-timeline metrics and OOD-All Event-IoU: Table 13.

Variant	ID								OOD-All					
	AUROC	FAR	Attr. F1	AV-mAP	A-mAP	V-mAP	A-IoU	V-IoU	AUROC	FAR	Attr. F1	AV-mAP	A-mAP	V-mAP
w/o intermediate-layer features	92.73 ± 0.17	4.93 ± 0.69	75.91 ± 0.52	84.75 ± 0.13	88.94 ± 0.12	80.56 ± 0.20	80.05 ± 0.71	65.24 ± 1.21	79.43 ± 0.25	5.29 ± 0.66	60.92 ± 0.69	53.73 ± 0.47	68.53 ± 0.67	42.66 ± 0.65
w/o interface features	<u>95.58</u> ± 0.16	6.42 ± 0.80	<u>83.35</u> ± 0.51	<u>90.61</u> ± 0.10	<u>92.12</u> ± 0.24	<u>89.10</u> ± 0.12	**84.81** ± 0.30	**79.76** ± 1.51	81.83 ± 0.24	8.84 ± 0.16	**64.82** ± 1.74	<u>58.38</u> ± 0.40	<u>72.95</u> ± 0.32	47.31 ± 0.76
w/o host conditioning	95.41 ± 0.08	5.00 ± 0.93	82.25 ± 0.48	89.86 ± 0.57	91.36 ± 0.86	88.36 ± 0.32	83.77 ± 0.33	74.13 ± 0.78	<u>81.90</u> ± 0.82	4.45 ± 1.15	62.24 ± 1.05	57.72 ± 1.55	69.24 ± 2.93	<u>47.89</u> ± 0.83
Full NFSR	**96.04** ± 0.10	3.52 ± 0.33	**83.61** ± 0.04	**91.30** ± 0.13	**92.24** ± 0.16	**90.36** ± 0.28	<u>84.72</u> ± 0.55	<u>76.90</u> ± 0.78	**84.61** ± 0.12	3.83 ± 0.44	<u>64.77</u> ± 0.72	**62.45** ± 0.08	**73.66** ± 0.79	**53.10** ± 0.56

leads in AUROC, attribution F1, and V-IoU on held-out TI2V, and the TASLE-trained models lead source-transfer V-mAP because those events come almost entirely from their training distribution. Without audio output, however, their AV-mAP is at most 37.50 in the two settings with audio edits.

False alarms are read together with coverage. Low FAR alone does not indicate reliable decisions, since a conservative detector can miss manipulations; we therefore read FAR together with attribution F1 and Event-IoU, which give no credit for missed events at the same operating point. Without recalibration, NFSR's realized FAR stays within 2.97–4.67% on all evaluation sets, close to the 5% budget. On source transfer, UMMAFormer's 12.21% FAR comes with lower coverage than NFSR on both tracks (Table 12). Visual coverage under shift remains the main open problem: NFSR reaches 26.42–41.21 V-IoU, and the two DAV-Det rows cover more visual events on source transfer but at 24.03% and 40.45% FAR. Corruption can likewise lower recall and FAR together (Appendix F.5).

5.3 Ablation Studies

We remove one component at a time under the same training and model-selection protocol (Table 4).

Each component matters more under distribution shift. Removing any component lowers OOD-All AV-mAP, with larger drops than in-domain. Intermediate-layer features contribute most: removing them reduces AV-mAP by 6.55 points on ID and 8.72 on OOD-All. Removing interface features has little effect on ID AV-mAP but reduces OOD-All AV-mAP by 4.07 points and more than doubles OOD FAR. Removing host conditioning lowers OOD-All AV-mAP by 4.73 points. The ablated variants also achieve higher Joint-mAP on ID but lower under shift (Table 13). Appendix D.3 further shows that both interface-predictable and residual intermediate-layer features remain discriminative.

Separated readouts preserve modality-specific decisions. On audio-only speech replacements (Appendix F.6), the separated visual readout fires on 2.9% of both edited videos and pristine twins, whereas the visual head of a cross-modal readout, whose branches exchange information through cross-modal attention, fires on 4.6% of pristine twins but 11.8–15.1% of the edited videos.

6 Conclusion

We introduce AV-FABLE, a dual-timeline benchmark for localizing sparse audio and visual manipulations in open-scene long-form videos. By preserving both temporal boundaries and modality identity, AV-FABLE exposes failures that joint-timeline evaluation can miss. Our experiments reveal a clear gap in current AV-LLMs: zero-shot prompting remains near chance, while supervised readouts of frozen perceptual representations recover strong forensic signals. Building on this observation, NFSR achieves the highest AV-mAP in-domain and across all three OOD settings, showing the effectiveness of modality-separated temporal readouts for long-form audio-visual forensics.

Limitation and future direction. AV-FABLE also opens several directions for future research. Extending the benchmark to broader real-world editing processes, refining perceptual boundary annotations, improving robustness to media corruption, and developing more faithful evidence-grounded explanations are promising next steps. We will publicly release the dataset, code, and model weights to support further work in long-form audio-visual forensics.

AI USE STATEMENT

Generative AI models were used in the SEE-Forge pipeline to construct AV-FABLE, including event-level edit planning, localization of source events, edit-feasibility checks, audio and visual media generation and editing, and drafting candidate evidence annotations (Section 3.2 and Appendix A). Temporal manipulation labels were derived from the executed edits. Two reviewers independently assessed edit validity and evidence support using the edited media, pristine counterparts, and edit instructions, with disagreements adjudicated by a third reviewer. Events without supported evidence records were retained for detection and localization but not explanation. AI tools also provided language assistance during manuscript preparation. The authors take responsibility for the final manuscript, results, and accompanying artifacts.

ETHICS STATEMENT

This work aims to advance research on audio-visual manipulation detection, localization, and forensic analysis in long-form videos. AV-FABLE builds on existing research datasets and introduces controlled manipulations using generative models. We follow the access and usage terms of the corresponding source datasets and do not redistribute source media, or excerpts thereof, when such redistribution is restricted by third-party licenses. The public release will instead focus on event-level annotations, source identifiers and timestamps, provenance metadata, independently releasable generated assets, and reconstruction utilities. Beyond long-form localization, these generated segments, produced by recent video, audio, and joint audio-visual generators across 27 editing configurations, can also serve as a standalone resource: labeled with generator, route, and operation and paired through source timestamps with their pristine counterparts, they support real-versus-generated classification, and those with human-verified evidence records additionally support evidence-grounded explanation. Given the dual-use nature of deepfake research, we do not provide end-to-end pipelines intended for impersonation, fraud, or other malicious uses, and we avoid intentionally constructing defamatory or deceptive content targeting real individuals. Human review is used solely for data and annotation quality control. Additional details on data licensing, human review, and release policies are provided in the Appendix.

REPRODUCIBILITY STATEMENT

To support reproduction, the anonymized supplementary material contains the complete AV-FABLE dataset manifest; the split, OOD-setting, and threshold-calibration lists; an independent evaluator with unit tests; and the model code with the training configuration. To preserve double-blind review, we have not publicly released our repository or dataset elsewhere. Upon acceptance, the dataset and model weights will be made openly available in accordance with the source-data terms described in the Ethics statement. Additional implementation and dataset construction details are provided in the Appendix.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13826–13836. IEEE, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Peijun Bao, Anwei Luo, Gang Pan, Alex C. Kot, and Xudong Jiang. ActivityForensics: A comprehensive benchmark for localizing manipulated activity in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42987–42996, 2026.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-NMS—improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5562–5570. IEEE, 2017.

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

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 (DICTA)*, 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*, pp. 7414–7423, 2024. doi: 10.1145/3664647.3680795.

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.

Songjun Cao, Yuqi Li, Yunpeng Luo, Jianjun Yin, and Long Ma. Leveraging large multimodal models for audio-video deepfake detection: A pilot study. *arXiv preprint arXiv:2602.23393*, 2026.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pp. 4978–4982, 2024. doi: 10.21437/Interspeech.2024-2016.

Chen Chen, Runze Li, Zejun Zhang, Pukun Zhao, Fanqing Zhou, Longxiang Wang, and Haojian Huang. Memory-anchored multimodal reasoning for explainable video forensics. *arXiv preprint arXiv:2508.14581*, 2025a.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: AI-generated video detection on million-scale GenVideo benchmark. *Science China Information Sciences*, 69(6):162103, 2026a. doi: 10.1007/s11432-024-4894-0.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning (ICML)*, pp. 5178–5193. PMLR, 2023.

Xuanjun Chen, Shih-Peng Cheng, Jiawei Du, Lin Zhang, Xiaoxiao Miao, Chung-Che Wang, Haibin Wu, Hung-yi Lee, and Jyh-Shing Roger Jang. Localizing audio-visual deepfakes via hierarchical boundary modeling. *arXiv preprint arXiv:2508.02000*, 2025b.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X^2-DFD: A framework for explainable and extendable deepfake detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93388–93435, 2025c. doi: 10.52202/085713-2809.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025d. doi: 10.18653/v1/2025.acl-long.313.

Zixuan Chen, Depeng Wang, Hao Lin, Li Luo, Ke Xu, Ya Guo, Huijia Zhu, Tanfeng Sun, and Xinghao Jiang. AVID: A benchmark for omni-modal audio-visual inconsistency understanding via agent-driven construction. *arXiv preprint arXiv:2604.13593*, 2026b.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, 2025. doi: 10.1109/CVPR52734.2025.02691.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807. IEEE, 2017.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other: Audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 439–447, 2020. doi: 10.1145/3394171.3413700.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Qiuhong Ke, Jianfei Cai, and Jun Zhu. Omni2Sound: Towards unified video-text-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1661–1671, 2026.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130(1):33–55, 2022. doi: 10.1007/s11263-021-01531-2.

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

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio Open. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yue Feng, Jingjing Li, Qijia Lu, Wei Ji, Jingrou Zhang, Fei Shen, Xiao Li, Yizhen Jia, Qiang Chen, Limin Wang, Wentong Li, and Jie Qin. Explainable forensics of manipulated segments in untrimmed long videos. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Joel Frank and Lea Schönherr. WaveFake: A data set to facilitate audio deepfake detection. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Google DeepMind. Gemini 3.7 Flash model card. `https://deepmind.google/models/model-cards/gemini-3-7-flash/`, 2026.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? Revealing common biases in generated image detection datasets. In *Computer Vision – ECCV 2024 Workshops*, pp. 80–95, 2024. doi: 10.1007/978-3-031-92089-9_6.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18685–18694, 2025. doi: 10.1109/CVPR52734.2025.01741.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4358–4367, 2021. doi: 10.1109/CVPR46437.2021.00434.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015.

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 (ICLR)*, 2022.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pp. 28831–28841, 2025. doi: 10.1109/CVPR52734.2025.02685.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. TangoFlux: Super fast and faithful text to audio generation with flow matching and CLAP-ranked preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2886–2895, 2020. doi: 10.1109/CVPR42600.2020.00296.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. doi: 10.1109/ICCV51701.2025.01597.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Christos Koutlis and Symeon Papadopoulos. AuViRe: Audio-visual speech representation reconstruction for deepfake temporal localization. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7168–7177. IEEE, 2026.

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 (ICLR)*, 2026.

Qingcao Li, Miao He, Liang Yi, Qing Wen, Yitao Zhang, Hongshuo Jin, Peng Cheng, Zhongjie Ba, Li Lu, and Kui Ren. Divide and conquer: Multimodal video deepfake detection via cross-modal fusion and localization. *arXiv preprint arXiv:2602.00209*, 2026a.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Xinze Li, Bingyu Zhu, Wuhui Duan, Congang Chen, Zeyu Fu, Yi Dong, et al. OMNI-Fake: Benchmarking unified multimodal social media deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30299–30311, 2026b.

Yaqi Li, Jielun Peng, Yabin Wang, Jincheng Liu, and Xiaopeng Hong. PATE-Forensics: Perception-as-tool for explainable deepfake forensics with general-purpose MLLMs. *arXiv preprint arXiv:2608.18573*, 2026c.

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. 3207–3216, 2020. doi: 10.1109/CVPR42600.2020.00327.

Zhenliang Li, Yutao Hu, Qixiong Wang, Wenpeng Du, Hongxiang Jiang, Jiasong Wu, Xiaolong Jiang, and Jungong Han. Impostor: An agent-curated benchmark for realistic AIGC manipulation localization. *arXiv preprint arXiv:2606.04545*, 2026d.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. ThinkSound: Chain-of-thought reasoning in multimodal LLMs for audio generation and editing. In *Advances in Neural Information Processing Systems*, volume 38, pp. 78526–78552, 2025. doi: 10.52202/085713-2630.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for AI-generated video detection. *arXiv preprint arXiv:2601.11035*, 2026.

Jiahao Mei, Heinrich Dinkel, Yadong Niu, Xingwei Sun, Gang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, Jian Luan, and Mengyue Wu. Dasheng AudioGen: A unified model for generating coherent audio scenes from text. *arXiv preprint arXiv:2605.27838*, 2026.

Changtao Miao, Yi Zhang, Weize Gao, Zhiya Tan, Weiwei Feng, Man Luo, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, et al. DDL: A large-scale datasets for deepfake detection and localization in diversified real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025.

MiniMax. MiniMax H3. `https://github.com/MiniMax-AI/MiniMax-H3`, 2026.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. GenVidBench: A 6-million benchmark for AI-generated video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15582–15590, 2026.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. OmniWeaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

Kyoungjun Park, Yifan Yang, Juheon Yi, Shicheng Zheng, Muhammad Muaz, Yifei Shen, Dongqi Han, Caihua Shan, and Lili Qiu. VidGuard-R1: AI-generated video detection and explanation via reasoning MLLMs and RL. In *International Conference on Learning Representations (ICLR)*, 2026.

Jielun Peng, Yabin Wang, Yaqi Li, Jincheng Liu, Xiaopeng Hong, and Athanasios V. Vasilakos. Less is more: Modality-decoupling for general AIGC audio-video detection. *arXiv preprint arXiv:2607.25543*, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518. PMLR, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Andreas Rössler, 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, 2019. doi: 10.1109/ICCV.2019.00009.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. HunyuanVideo-Foley: Multimodal diffusion with representation alignment for high-fidelity Foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shen Shen, Quan Zhang, Dan Jiang, and Ke Zhang. EVAS: Efficient multimodal temporal forgery localization via audio-visual synergy and steered boundary calibration. *arXiv preprint arXiv:2607.04472*, 2026.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18815–18825, 2025. doi: 10.1109/CVPR52734.2025.01753.

Hao Tan, Jun Lan, Senyuan Shi, Zichang Tan, Zijian Yu, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. VideoVeritas: AI-generated video detection via perception pretext reinforcement learning. *arXiv preprint arXiv:2602.08828*, 2026a.

Hao Tan, Jun Lan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations (ICLR)*, 2026b.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-Enhanced Audio-Visual Large Language Models. *arXiv preprint arXiv:2506.15220*, 2025.

Shahroz Tariq, David Nguyen, M. A. P. Chamikara, Tingmin Wu, Alsharif Abuadbba, and Kristen Moore. LLMs are not yet ready for deepfake image detection. *arXiv preprint arXiv:2506.10474*, 2025.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: A unified framework for anything-to-audio generation. In *International Conference on Learning Representations*, volume 2026, pp. 110625–110651, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, et al. DFBench: Benchmarking deepfake image detection capability of large multimodal models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12666–12673, 2025. doi: 10.1145/3746027.3758204.

Jiarui Wang, Huiyu Duan, Juntong Wang, and Xiongkuo Min. FVBench: Benchmarking deepfake video detection capability of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4425–4437, 2026.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020. doi: 10.1016/j.csl.2020.101114.

Junyan Wu, Wei Lu, Xiangyang Luo, Rui Yang, Qian Wang, and Xiaochun Cao. Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7395–7403, 2024. doi: 10.1145/3664647.3680585.

Shuhan Xia, Peipei Li, Xuannan Liu, Dongsen Zhang, Xinyu Guo, and Zekun Li. AVFakeBench: A comprehensive audio-video forgery detection benchmark for AV-LMMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35416–35426, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations (ICLR)*, 2025c.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 47–54, 2021. doi: 10.21437/ASVSPOOF.2021-8.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4534–4565, 2023. doi: 10.52202/075280-0201.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. DF40: Toward next-generation deepfake detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29387–29434, 2024. doi: 10.52202/079017-0925.

Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. DyFADet: Dynamic feature aggregation for temporal action detection. In *European Conference on Computer Vision (ECCV)*, pp. 305–322, 2024. doi: 10.1007/978-3-031-72952-2_18.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. Half-truth: A partially fake audio detection dataset. In *Proc. Interspeech*, pp. 1654–1658, 2021. doi: 10.21437/Interspeech.2021-930.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, et al. ADD 2023: The second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*, 2023.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Chu Yuan Zhang, Cunhang Fan, Zhengkun Tian, Haoxin Ma, and Ruibo Fu. SceneFake: An initial dataset and benchmarks for scene fake audio detection. *Pattern Recognition*, 152:110468, 2024. doi: 10.1016/j.patcog.2024.110468.

Qilin Yin, Wei Lu, Xiangyang Luo, Rui Yang, and Xiaochun Cao. Context-aware TFL: A universal context-aware contrastive learning framework for temporal forgery localization. *IEEE Transactions on Dependable and Secure Computing*, 2026.

Tao Yu, Yujia Yang, Shenghua Chai, Zhang Jinshuai, Haopeng Jin, Hao Wang, Minghui Zhang, Zhongtian Luo, Yuchen Long, Xinlong Chen, et al. When seeing is not believing–a benchmark for search-grounded video misinformation detection. *arXiv preprint arXiv:2606.04098*, 2026.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *European Conference on Computer Vision (ECCV)*, pp. 492–510, 2022. doi: 10.1007/978-3-031-19772-7_29.

Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, and Junichi Yamagishi. The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:813–825, 2023a. doi: 10.1109/TASLP.2022.3233236.

Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. UMMAFormer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8749–8759, 2023b. doi: 10.1145/3581783.3613767.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026. doi: 10.1007/s11263-025-02649-3.

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

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12993–13000, 2020. doi: 10.1609/aaai.v34i07.6999.

Jiafeng Zhong, Bin Li, and Jiangyan Yi. Enhancing partially spoofed audio localization with boundary-aware attention mechanism. In *Proc. Interspeech*, pp. 4838–4842, 2024. doi: 10.21437/Interspeech.2024-587.

Xiaodong Zhu, Suting Wang, Junqi Yang, Yuhong Yang, Weiping Tu, and Zhongyuan Wang. Query-based audio-visual temporal forgery localization with register-enhanced representation learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8547–8556, 2025. doi: 10.1145/3746027.3755563.

APPENDIX CONTENTS

A AV-FABLE: Construction and Data Quality **18**
- A.1 Sources, Composition, and Labels . . . 18
- A.2 Event-Conditioned Construction . . . 18
- A.3 Speech Replacement . . . 19
- A.4 Human Review and Decision Agreement . . . 20
- A.5 Metadata-Only Shortcut Probes . . . 20
- A.6 Extended Benchmark Comparison . . . 21

B Extended Related Work **22**
- B.1 Benchmark scope and temporal supervision . . . 22
- B.2 Generation and localized editing . . . 24
- B.3 Temporal localization and forensic representations . . . 24

C Evaluation and Implementation **25**
- C.1 Splits and OOD Evaluation Protocol . . . 25
- C.2 Scoring and Operating Points . . . 26
- C.3 Native-Track and Joint-Timeline Evaluation . . . 26
- C.4 Readout and Baselines . . . 28

D Forensic Evidence in Frozen Representations **29**
- D.1 Probe Design and Layer Selection . . . 29
- D.2 Encoder Replacement . . . 30
- D.3 Complementary Evidence across Representation Stages . . . 31

E Sparse Evidence in Long Videos **32**

F Generalization and Sensitivity Analyses **32**
- F.1 Seed Stability . . . 32
- F.2 Value of AV-FABLE Supervision . . . 32
- F.3 Training-Set Scale . . . 33
- F.4 External Comparison with Official MSLoc . . . 33
- F.5 Robustness to Input Degradation . . . 34
- F.6 Cross-Track False Alarms . . . 34

G Language Verification and Temporal Decisions **35**
- G.1 Zero-Shot and Direct-Transfer Results . . . 35
- G.2 Fixed-Candidate Verification . . . 35
- G.3 Language Rescoring and Score Fusion . . . 36

H Explanation Baselines **36**
- H.1 Training and Evaluation Setup . . . 36
- H.2 Structured Output and Semantic Identification . . . 37
- H.3 Media Dependence of Support Decisions . . . 37

I Qualitative Examples **38**



Figure 3: Scene diversity in AV-FABLE. Example frames span indoor and outdoor environments, human activities, objects, food preparation, sports, and other contexts represented in the twin pairs.

A AV-FABLE: Construction and Data Quality

A.1 Sources, Composition, and Labels

ActivityNet-Raw supplies open-scene third-person videos, while EPIC-KITCHENS-100 supplies egocentric recordings (Heilbron et al., 2015; Damen et al., 2022). The EPIC construction pool contains 400 source recordings from 37 participants, totaling 67.3 hours; natural event clips are extracted before local manipulation and matched export. HDTF provides held-out talking-head hosts (Zhang et al., 2021). Source qualification checks decodability, both media tracks, synchronization, usable context, and observable candidate events.

The release comprises 16,650 twin pairs and 22,153 track-specific events (Table 5). The total of 998.1 hours sums both members of every pair; counting each pair once gives 499.1 hours. Video-level duration and coverage distributions count each pair once, whereas event statistics count the annotated intervals on each modality separately. The 16,650 manipulated videos contain 4,903 audio-only, 6,376 visual-only, and 5,371 both-track examples. Independent timelines allow events to overlap or occur asynchronously; jointly generated media can yield separate audio and visual event annotations. Event records include identity, track, interval, operation, target type, execution route, generator, and evidence status. Operations comprise replacement (14,914 events), insertion (5,555), regeneration (961), and removal (723). Targets cover sound events, actions, objects, attributes, ambience, speech, scene regions, text, and background.

A.2 Event-Conditioned Construction

SEE-Forge first forms an event catalog from captions/ASR, audio onsets, motion, shot boundaries, and interactions. A timeline localizer verifies natural event boundaries. The planner selects an existing event and specifies its modality, interval, single-action edit instruction, operation, target, preservation constraints, execution route, and expected evidence. Gemini 3.7 Flash plans (Google DeepMind, 2026), Qwen2.5-VL localizes (Bai et al., 2025), and Qwen3-Omni reviews plan validity and executability (Xu et al., 2025b). After execution, Gemini 3.7 Flash also drafts candidate evidence records from a clip around each event. Checks establish that the target is observable, the requested change is compatible with its context, and the selected generator can execute the plan.

Routing matches the edit to a generator: masked V2V for local object/attribute changes, V2V for process changes, frame-conditioned generation for endpoint-constrained events, joint generation for

Table 5: Release statistics. Counts are per pair or annotated event as indicated; multi-event rate is over manipulated videos. External TASLE and supplementary TTS examples are excluded.

Quantity	Train	Val.	ID test	OOD
Twin pairs	10,954	2,347	2,347	1,002
Audio events	6,843	1,451	1,482	530
Visual events	7,911	1,729	1,695	512
Video duration: median (s)	88	88	91	87
Video duration: P10–P90 (s)	36–204	33–205	36–207	17–238
Event duration: median (s)	9.4	9.3	9.5	5.7
Event duration: P10–P90 (s)	4.9–12.7	5.0–12.6	5.0–12.9	4.5–10.3
Multi-event rate	34.2%	35.0%	35.0%	2.8%
Events with candidate evidence	13,628	2,948	2,951	1,002

audio-visual changes, and Foley/text-to-audio/TTS for audio edits. Generators receive the target interval and 2–5 seconds of context. Reinsertion preserves non-target content, checks visual anchors and motion continuity, and matches audio loudness with short cross-fades. Persistent entity changes require consistent editing across their visible intervals or replanning. Diversity memory controls repeated scene, target, operation, and generator combinations.

Both twins then pass through a common export profile, matching container, codec, resolution, frame rate, sample rate, channel count, and encoding policy. Pristine twins follow the same cropping, reinsertion, and export steps while retaining the original target content. Figure 1 summarizes the pipeline; Figures 4 and 5 provide the dataset statistics and detailed construction coverage.

Table 6: Generator coverage across the 23 training and four OOD-only editing configurations. Each configuration pairs a generator with an execution route.

Generator group	Routes	Role
MiniMax H3 (MiniMax, 2026); LTX 2.3/2.5 (HaCohen et al., 2025)	FLF2V, FLF2VA, reference / retake	Train / ID
Wan Fun / VACE (Wan et al., 2025; Jiang et al., 2025); CogVideoX-Fun (Yang et al., 2025)	Frame-conditioned, masked V2V, V2V	Train / ID
MMAudio (Cheng et al., 2025); FoleyCrafter (Zhang et al., 2026); HunyuanVideo-Foley (Shan et al., 2025); ThinkSound (Liu et al., 2025)	Video-to-audio Foley	Train / ID
Stable Audio (Evans et al., 2025); TangoFlux (Hung et al., 2026); AudioX (Tian et al., 2026); Dasheng AudioGen (Mei et al., 2026)	Text-to-audio	Train / ID
Qwen3-TTS; F5-TTS v1 (Hu et al., 2026; Chen et al., 2025d)	Speech replacement	Train / ID
Omni2Sound (Dai et al., 2026); Wan-FLF2V (Wan et al., 2025); OmniWeaving (Pan et al., 2026)	Foley, FLF2V, V2V	OOD generator
Wan2.2-TI2V-5B (Wan et al., 2025)	TI2V	OOD route + generator

A.3 Speech Replacement

The speech-replacement subset contains 518 ActivityNet examples (351/77/90 in train/validation/ID test), each combining a speech replacement with a visual manipulation. Eligible hosts contain 5–12 seconds of continuous English speech and a separate same-speaker reference. Qwen3-TTS and F5-TTS produce 258 and 260 replacements, using unrelated transcripts of similar length (Hu et al., 2026; Chen et al., 2025d). Vocal separation preserves background audio; loudness matching and 20-ms cross-fades preserve the timeline. The median replacement lasts 7.3 seconds, and these examples carry both-track attribution labels.

Table 7: Initial agreement on retention and explanation eligibility, and the overall third-reviewer adjudication rate. Rates use the complete reviewed event pool as the denominator.

Metric	Value
Retention: raw agreement	96.4%
Retention: Cohen's κ	0.841
Explanation eligibility: Cohen's κ	0.823
Overall third-reviewer adjudication rate	7.7%

A separate audio-only supplement uses 242 pristine ID hosts with matched interval/text budgets: 239 Qwen3-TTS, 239 F5-TTS, and 238 XTTS-v2 examples (Casanova et al., 2024). The 90 ID-test speech replacements are built on these hosts and reuse their generated speech. The language-verification analyses (Appendix G) use an earlier training release without speech replacement.

A.4 Human Review and Decision Agreement

Two reviewers independently assessed all 22,153 events using the edited media, pristine twins, and edit instructions. Retention (retain vs. reject/rework) required the intended edit to be realized while preserving non-target content and avoiding unintended integration failures. This decision also covered boundary quality: events whose perceptible change clearly mismatched the recorded interval, or whose interval boundaries showed abrupt transitions or visual breakdown, were rejected or reworked. Reviewers did not re-draw boundaries, so retained intervals remain the executed edit intervals. Explanation eligibility required candidate evidence statements to be supported by the media; otherwise valid events remained usable for detection/localization only. Table 7 reports initial agreement on these two binary decisions before adjudication, together with the overall rate of referral to a third reviewer.

A.5 Metadata-Only Shortcut Probes

We test whether file-level properties distinguish pristine and manipulated videos after matched export. Logistic regression and gradient-boosted trees receive container format, codec, resolution, frame rate, timebase, bitrate, audio channels and sample rate, frame count, duration, and encoder tags, without access to decoded frames or waveforms. Five-fold evaluation keeps each twin pair in the same fold. The multifield probes remain close to chance on both ID and the OOD pool (Table 8), and

Table 8: Metadata-only authenticity probes: AUROC (%). Ranges summarize the evaluated multifield configurations. The OOD pool, used throughout the appendix analyses, pools all 1,002 AV-FABLE OOD pairs across the source, generator, and TI2V settings, without TASLE or equal weighting of settings.

Control	ID	OOD pool
Multifield probes: AUROC range	53.9–54.6	51.0–53.0
Strongest single field	51.9	–

the strongest single field reaches only 51.9 ID AUROC. File-level properties therefore carry little label information after matched export. The probes do not inspect decoded frames or waveforms, however, so they bound only this class of shortcut.

Controlled and remaining shortcut risks. By construction, each twin pair shares its host content, and both twins undergo the same cropping, reinsertion, and export steps; source-disjoint splits prevent shared-source leakage; and the metadata probes show that file-level format carries little label information. These controls do not exclude decoded-signal traces that generators may introduce inside the edited interval beyond the intended semantic change, such as local resampling by generators that operate at a different resolution, frame rate, or sample rate, generator-specific texture, and acoustic residue from synthesis. Some of these traces also accompany real manipulations and are legitimate forensic evidence, whereas others may be specific to the generators used here. The held-out-generator setting measures transfer to unseen generators on seen routes but does not separate these two cases, so pipeline-specific cues cannot be fully excluded.

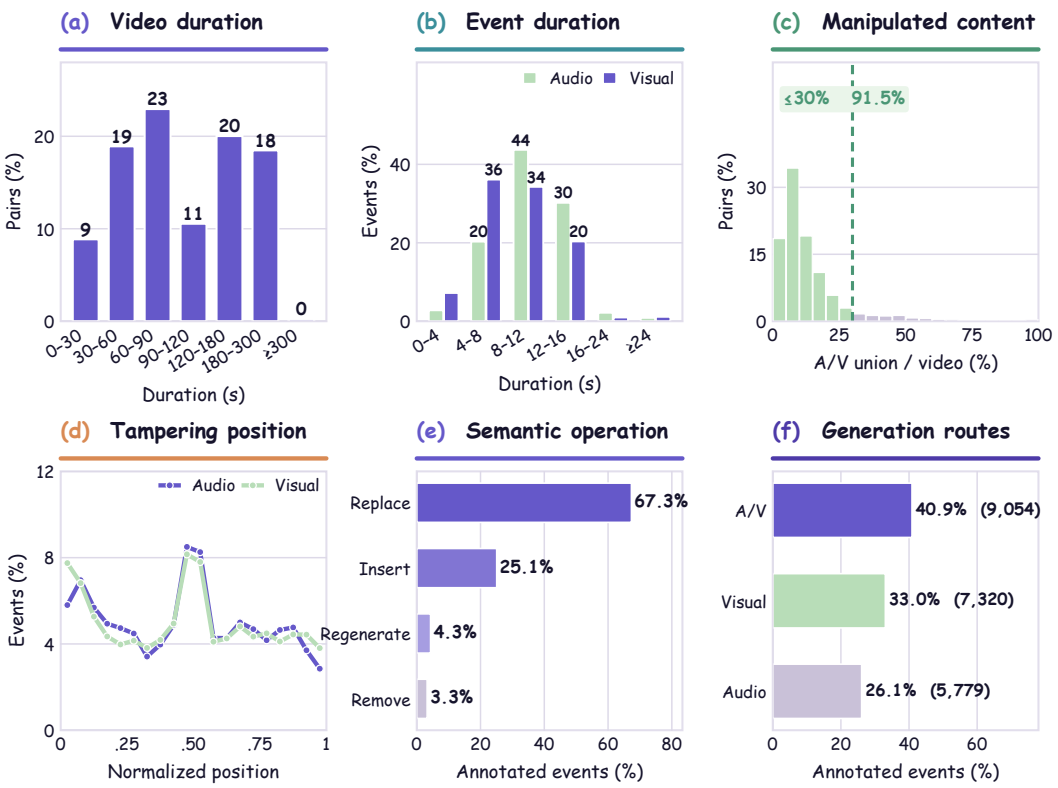


Figure 4: Statistics and manipulation coverage in AV-FABLE. The six panels show video duration, event duration by modality, manipulated temporal coverage, tampering position, semantic operation, and aggregated generation route. Video-level distributions count each twin pair once; coverage is computed over the union of audio and visual intervals.

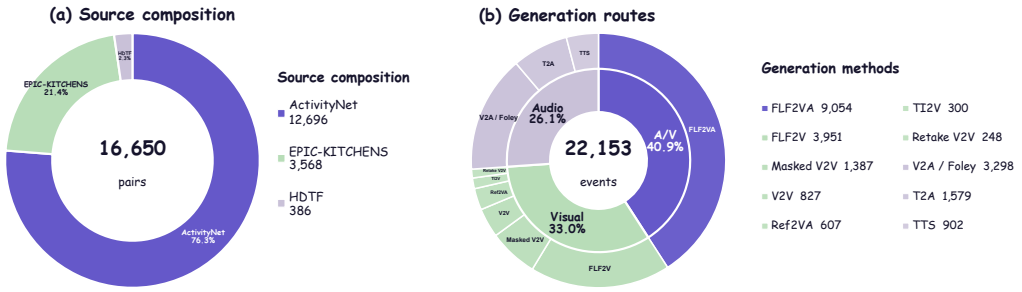


Figure 5: Source and generation composition in AV-FABLE. The left donut shows source-domain shares; the right donut shows modality attribution and the aggregated generation methods used to produce the edited events. The full list of editing configurations is given in Table 6.

A.6 Extended Benchmark Comparison

Table 9 complements the main-text comparison with a broader view of dataset scale, temporal structure, annotation coverage, and evaluation design.

Table 9: **Extended comparison with forensic datasets.** Additional video, audio, and image benchmarks. Columns and symbols follow Table 1.

Dataset	Scale	Scene	Manip.		Loc.	Expl.	Pair	OOD shift		
			A	V				Src	Gen	Type/ route
Temporal and audio–visual benchmarks										
AV-Deepfake1M (Cai et al., 2024)	1,146,760 v	Talking heads	✓	✓	✓	×	(✓)	×	×	×
DDL-AV (Miao et al., 2025)	330K v	Face-centric AV	✓	✓	✓	×	(✓)	(✓)	(✓)	×
ArEnAV (Kuckreja et al., 2026)	387,072 v	Arabic/English speech	✓	✓	✓	×	(✓)	×	×	×
X-AVFake (FakeHunter) (Chen et al., 2025a)	$> 5,700$ v	Open AV scenes	✓	✓	(✓)	✓	(✓)	×	×	×
AVFakeBench (Xia et al., 2026)	3,000 v 12,000 QA	Human + general	✓	✓	×	✓	×	×	×	×
Omni-Fake (Li et al., 2026b)	∼1.35M items	I / A / V / AV-TH	✓	✓	(✓)	✓	×	(✓)	(✓)	(✓)
Video datasets and evaluation suites										
DeeperForensics-1.0 (Jiang et al., 2020)	60,000 v	Faces + perturb.	×	✓	×	×	(✓)	×	×	×
DF40 (Yan et al., 2024)	> 0.1M v > 1M img	Faces; 40 methods	×	✓	×	×	(✓)	✓	✓	(✓)
GenVideo (DeMamba) (Chen et al., 2026a)	2,314,182 v	Open scenes	×	✓	×	×	×	(✓)	(✓)	×
GenVidBench (Ni et al., 2026)	6,784,490 v	Open scenes	×	✓	×	×	(✓)	(✓)	✓	×
AIGVDBench (Ma et al., 2026)	> 440K v	Open scenes	×	✓	×	×	(✓)	×	✓	×
VidGuard-R1 data (Park et al., 2026)	140,000 v	Open scenes	×	✓	×	(✓)	(✓)	(✓)	(✓)	×
MintVid (Tan et al., 2026a)	4,992 v	General / face / fact	×	✓	×	×	(✓)	(✓)	(✓)	×
MVAD / DDL-GAV (reported) (Peng et al., 2026)	205,758 AV	Realistic + anime	✓	✓	×	×	×	×	×	×
EVID-Bench (Yu et al., 2026)	222 v	Web misinformation	×	✓	×	✓	(✓)	×	×	×
DeepfakeBench (v2) (Yan et al., 2023)	9 base sets + DF40	Benchmark suite	×	✓	×	×	(✓)	(✓)	(✓)	×
FVBench (released subset) (Wang et al., 2026)	8,398 rows	Public video subset	×	✓	×	×	×	(✓)	✓	(✓)
Audio-only forensic datasets										
PartialSpoof (Zhang et al., 2023a)	121,461 utt	English speech	✓	×	✓	×	(✓)	×	✓	×
HAD (partial track) (Yi et al., 2021)	107,224 trials	Mandarin speech	✓	×	✓	×	(✓)	×	✓	×
ADD 2023 (Track 2) (Yi et al., 2023)	120,917 utt	Mandarin speech	✓	×	✓	×	(✓)	×	(✓)	×
ASVspoof 2019 (Wang et al., 2020)	121,461 LA trials 218,430 PA trials	Speech / replay	✓	×	×	×	(✓)	(✓)	✓	×
ASVspoof 2021 (Yamagishi et al., 2021)	181,566 LA trials 943,110 PA trials 611,829 DF trials	Speech / replay	✓	×	×	×	(✓)	(✓)	(✓)	×
SceneFake (Yi et al., 2024)	84,480 trials	Speech + soundscape	✓	×	×	×	(✓)	(✓)	(✓)	×
WaveFake (Frank & Schönherr, 2021)	117,985 fake utt	English / Japanese	✓	×	×	×	(✓)	(✓)	✓	×
Image / explanation resources										
DDL-I (Miao et al., 2025)	1.433M img	Faces / varied images	×	✓	×	×	(✓)	(✓)	(✓)	×
SID-Set (SIDA) (Huang et al., 2025)	300K img 3K explanations	Open images	×	✓	×	(✓)	(✓)	×	(✓)	×
MMTD-Set-34k (FakeShield) (Xu et al., 2025c)	34,470 ann.	PS / face / AIGC edit	×	✓	×	✓	(✓)	(✓)	(✓)	×
HydraFake (Veritas) (Tan et al., 2026b)	100,000 img	Faces; multi-forgery	×	✓	×	(✓)	(✓)	(✓)	✓	(✓)
Impostor (Li et al., 2026d)	100K fake img	Open image edits	×	✓	×	×	(✓)	(✓)	(✓)	×
DFBench (Wang et al., 2025)	540K img	Open images	×	✓	×	×	(✓)	×	(✓)	×
X^2-DFD annotation set (Chen et al., 2025c)	–	Faces / VQA	×	✓	×	✓	(✓)	(✓)	(✓)	×
Tariq et al. evaluation (Tariq et al., 2025)	100 img	Portrait images	×	✓	×	×	×	×	×	×

B Extended Related Work

B.1 Benchmark scope and temporal supervision

From facial authenticity to general video forensics. Visual deepfake evaluation has largely developed around facial authenticity, as in FaceForensics++, Celeb-DF, and DFDC (Rössler et al., 2019; Li et al., 2020; Dolhansky et al., 2020). DeeperForensics-1.0 and DF40 expand this setting through challenging perturbations and a broader range of manipulation techniques (Jiang et al., 2020; Yan et al., 2024). Temporal supervision also appears in ForgeryNet, alongside image- and video-level face-forgery tasks (He et al., 2021). General generated-video resources such as GenVideo, GenVidBench, and AIGVDBench broaden the content and generators considered (Chen et al., 2026a; Ni et al., 2026; Ma et al., 2026). Moving from wholly generated clips to sparse edits within a video requires both broad content coverage and interval annotations.

Partial speech manipulation and acoustic scenes. Speech anti-spoofing provides a parallel route from recording-level decisions to within-recording localization. ASVspoof 2019 and ASVspoof 2021 establish protocols covering synthesized, converted, replayed, and deepfake speech (Wang et al., 2020; Yamagishi et al., 2021); WaveFake supplies synthetic speech from multiple generation methods (Frank & Schönherr, 2021). PartialSpoof, HAD, and ADD 2023 Track 2 focus more directly on partial manipulation or the localization of manipulated regions (Zhang et al., 2023a; Yi et al., 2021; 2023). Here, authentic and manipulated content can coexist within one recording, making global authenticity labels insufficient for boundary estimation. SceneFake broadens the manipulated

content from speech to the surrounding acoustic scene (Yi et al., 2024). This distinction is relevant to environmental-sound edits, whose forensic evidence need not involve a speaker or lip synchronization.

Audio-visual benchmarks beyond talking heads. FakeAVCeleb combines facial and speech manipulations, establishing a multimodal detection setting (Khalid et al., 2021). Within speaker-centered video, LAV-DF and AV-Deepfake1M provide content-driven edits with temporal annotations (Cai et al., 2022; 2024). AV-Deepfake1M++ extends this line with larger-scale evaluation under realistic perturbations (Cai et al., 2025). DDL-AV also retains fine-grained localization annotations across audio, visual, and joint manipulations (Miao et al., 2025), while ArEnAV addresses Arabic–English code-switching and dialect variation (Kuckreja et al., 2026). More general audio-video content is represented by DDL-GAV, built on MVAD and described in DAV-Det's evaluation (Peng et al., 2026). The emphasis shifts toward forensic interpretation in X-AVFake and AVFakeBench, which include explanatory annotations or question-based evaluation beyond conventional talking-head detection (Chen et al., 2025a; Xia et al., 2026). Omni-Fake brings several media types into one benchmark, but its individual-modality localization tasks differ from its binary audio-video talking-head task (Li et al., 2026b). These benchmarks vary not only in content coverage but also in whether their labels supervise clip-level decisions or modality-specific intervals.

Long-form context and modality-specific targets. Sparse edits within extended authentic context are central to TASLE and ActivityForensics, which study visual manipulation localization in untrimmed or activity-rich videos (Feng et al., 2026; Bao et al., 2026). AVID brings long-form context to audio-visual inconsistency understanding, including temporal grounding and reasoning about conflicting streams (Chen et al., 2026b). This shares the need to identify brief events, but an inconsistency target is not the same as establishing manipulation in each modality. A replacement can remain semantically compatible with the other stream, and a mismatch need not determine which stream was altered. AV-FABLE instead assigns executed edits to separate audio and visual timelines, including asynchronous events within the same host video. Native-track evaluation matches each prediction against its own modality, so a temporally correct prediction on the wrong stream receives no credit.

Reasoning resources and evaluation suites. Language-based forensic evaluation also depends on what evidence a model can access. The datasets accompanying VidGuard-R1 and VideoVeritas, including the latter's MintVid, support reasoning-oriented generated-video detection (Park et al., 2026; Tan et al., 2026a). EVID-Bench requires web retrieval and cross-video comparison to verify video misinformation, including cases not resolvable from the input video alone (Yu et al., 2026). It therefore differs from detecting local synthesis traces, even when AI-generated content contributes to the misinformation. DeepfakeBench standardizes detector implementation and evaluation (Yan et al., 2023), whereas FVBench assesses the deepfake-video detection capabilities of large multimodal models (Wang et al., 2026). The extended comparison in Table 9 includes these evaluation resources alongside datasets without assuming that they share a localization protocol.

Spatial localization and explanation precedents. Image forensics provides relevant precedents for localized evidence, although its annotations refer to spatial regions rather than temporal intervals. DDL-I supplies image-forgery data and spatial annotations, complementing the DDL-AV branch (Miao et al., 2025). SID-Set and MMTD-Set-34k support SIDA and FakeShield, respectively, linking image detection and localization to explanatory supervision (Huang et al., 2025; Xu et al., 2025c). HydraFake, introduced with Veritas, emphasizes generalization across forgery techniques and domains (Tan et al., 2026b); Impostor addresses realistic localized AI image editing (Li et al., 2026d). DFBench and the evaluation by Tariq et al. assess multimodal models on deepfake-image detection (Wang et al., 2025; Tariq et al., 2025). The X^2-DFD annotation set further supports explainable deepfake detection (Chen et al., 2025c). They provide precedents for coupling localized evidence with explanation, while the temporal dimension requires a different annotation target.

Construction controls and generalization. The forensic task can be distorted when incidental processing traces predict the label. Circumventing Shortcuts identifies exploitable artifacts in audio-visual datasets (Smeu et al., 2025), while Fake or JPEG? and B-Free examine preprocessing and training-data biases in visual detection (Grommelt et al., 2024; Guillaro et al., 2025). These findings motivate matched processing of pristine and manipulated examples without implying that every

synthesis trace should be removed. Likewise, a held-out generator tests a different shift from a new source domain or editing route. When several factors change together, transfer measures the combined setting rather than an isolated causal effect of any one factor.

B.2 Generation and localized editing

From generative capability to a forensic event. Video models such as CogVideoX, LTX-Video, and Wan provide synthesis capabilities, while VACE also supports controlled video editing (Yang et al., 2025; HaCohen et al., 2025; Wan et al., 2025; Jiang et al., 2025). Speech can be replaced using synthesis systems such as F5-TTS and Qwen3-TTS (Chen et al., 2025d; Hu et al., 2026). For non-speech audio, MMAudio and FoleyCrafter condition generation on visual content (Cheng et al., 2025; Zhang et al., 2026), allowing an altered soundtrack to remain compatible with the visible scene. The same generator may produce a fully synthetic clip or a brief replacement inserted into an authentic host. The latter changes the forensic problem by retaining surrounding context and requiring event boundaries. In AV-FABLE, the executed edit coordinates define temporal ground truth, while the evidence records describe observable changes within and around those intervals.

B.3 Temporal localization and forensic representations

From classification scores to event intervals. Single-modality detection provides useful references but does not always train an explicit boundary predictor. Xception supplies a visual classification backbone (Chollet, 2017), AASIST targets speech anti-spoofing (Jung et al., 2022), and DeMamba models spatiotemporal evidence in generated videos (Chen et al., 2026a). Applying these detectors to successive windows produces local scores, but their conversion into intervals remains an additional modeling choice. A correct video-level decision can therefore coexist with inaccurate boundaries or predictions assigned to the wrong modality. For partially spoofed audio, CFPRF combines frame-level predictions with proposal refinement (Wu et al., 2024); boundary-aware attention explicitly models genuine–spoofed transitions (Zhong et al., 2024). These methods are closer to the interval-prediction objective than recording-level classification alone.

Cross-modal evidence and temporal context. Audio-visual localization can use relationships between streams as forensic evidence, as in dissonance-based detection (Chugh et al., 2020). Cross-modal reconstruction and synchronization offer related cues in AuViRe and ICS-Sync (Koutlis & Papadopoulos, 2026; Anshul et al., 2025). Such cues must still be converted into temporal predictions: BA-TFD and BA-TFD+ emphasize boundary learning (Cai et al., 2022; 2023), and HBMNet combines frame-level and proposal-level boundary information (Chen et al., 2025b). UMMAFormer and RegQAV develop adaptive multimodal representations and query-based predictions, respectively (Zhang et al., 2023b; Zhu et al., 2025). UniCaCLF relates genuine and forged instants to global context through contrastive learning (Yin et al., 2026), while EVAS combines audio-visual interaction with boundary calibration (Shen et al., 2026). Together, these approaches connect evidence aggregation to boundary prediction, establishing context modeling as a relevant precedent for long-form localization.

Temporal decoding and host conditioning. Temporal action detection provides architectures that predict event intervals from feature sequences, including ActionFormer and DyFADet (Zhang et al., 2022; Yang et al., 2024). Their relevance is architectural: recognizing an action and deciding which part of it was manipulated remain different supervision targets. Action boundaries need not coincide with manipulation boundaries, even when the same temporal decoding architecture is used. Feature-wise linear modulation provides a general mechanism for conditioning a representation on contextual information (Perez et al., 2018). NFSR uses the input host video's feature statistics to modulate local evidence before modality-specific interval prediction. The pristine twin supports benchmark construction and annotation, but it is not required as a reference at inference.

Modality-specific decisions. Cross-modal evidence is informative but does not establish that both streams have been altered. DAV-Det models audio and visual forensic evidence independently (Peng et al., 2026), and Divide and Conquer combines modality-specific detectors and localizers through score fusion (Li et al., 2026a). Decoupled video-level detection and separate temporal predictions

address different output requirements. The latter accommodates events with different boundaries in each stream, making modality identity part of the localization target.

Frozen representations and language interfaces. Representation quality is a different question from the behavior of a language-based prediction interface. Linear probes assess information accessible through restricted supervised readouts (Alain & Bengio, 2016); self-supervised deepfake studies examine the complementarity of audio and visual features (Boldisor et al., 2026). Qwen2.5-Omni, Qwen3-Omni, and video-SALMONN offer audio-visual language interfaces (Xu et al., 2025a;b; Tang et al., 2025), while AV-LMMDetect adapts an AV-LLM for authenticity classification (Cao et al., 2026). These settings differ in supervision, so weak zero-shot performance alone cannot establish that the underlying features lack useful forensic information. Section 5.1 compares temporal localizers on the same frozen features, separating the choice of encoder from the design of its readout.

Language-assisted localization and explanation. A language model may revise intervals or explain them, and these roles should be distinguished. TASLE's MSLoc uses language-assisted proposal refinement as part of visual localization (Feng et al., 2026). PATE-Forensics separates forensic perception from explanation generation (Li et al., 2026c); FakeHunter retrieves contextual evidence and invokes tools to support multimodal reasoning (Chen et al., 2025a). Image-level systems likewise connect detected spatial regions to natural-language explanations (Huang et al., 2025; Xu et al., 2025c; Chen et al., 2025c). In the separate explanation evaluation of Appendix H, language outputs describe fixed events and do not alter localization or detection decisions. Fluent descriptions do not themselves establish that predicted intervals or modality labels are correct, motivating separate evaluation of these outputs.

C Evaluation and Implementation

C.1 Splits and OOD Evaluation Protocol

All twins and derived samples from the same host video remain in one split. Training, validation, and ID testing use ActivityNet and EPIC-KITCHENS with shared manipulation categories and editing configurations. The three OOD settings are defined by their held-out components. Source transfer uses HDTF and a source-disjoint TASLE subset. Held-out generators retain editing routes represented in training while using unseen generators. Held-out TI2V jointly withholds the text-and-image-conditioned video generation route and its generator, Wan2.2-TI2V-5B. This setting evaluates transfer to that route–generator configuration, without separating their contributions to the distribution shift. The calibration subset contains 256 pristine and 256 manipulated videos from the source-validation set used for checkpoint selection. Thresholds are estimated from its pristine videos only; test and OOD data are used solely for evaluation. The source-transfer setting combines different domains

Table 10: OOD evaluation sets and their held-out components. Source-transfer detection and Joint-timeline results average HDTF and TASLE with equal domain weight, whereas native A-mAP uses HDTF and native V-mAP pools both domains; held-out-generator and TI2V results are reported on their respective sets.

Setting / domain	Manipulated	Pristine	Shift
Source / HDTF	386	386	New domain; predominantly audio edits
Source / TASLE	358	360	External dataset; visual edits
Generator / three models	311	311	Seen routes, new generators
TI2V / Wan2.2	300	300	Held-out route and generator

and modality compositions: HDTF contains 384 audio and two visual manipulated examples, while TASLE supplies visual forgeries. The TASLE subset contains 718 videos after excluding 54 examples from 15 EPIC source recordings shared with AV-FABLE training, using source metadata to determine membership. The held-out-generator set contains 311 of 317 candidate pairs after deduplication and restriction to common prediction coverage. TASLE is external to the AV-FABLE release counts. Equal weighting of HDTF and TASLE prevents the larger source subset from dominating the source-transfer result. Attribution macro-F1 is computed on the pooled 1,490 source videos instead, because each domain alone lacks some attribution classes; the pooled set contains clean, audio-only, and

visual-only videos but no both-track videos (Eq. 5). Native localization cannot use equal domain weighting either: HDTF contains all 384 audio events of the setting but only two visual events. Source-transfer A-mAP is therefore computed on HDTF, and V-mAP on the 393 pooled visual events of both domains. The resulting source-transfer scores reflect both domain and modality composition. Because none of the three settings isolates a single shift factor, differences between their scores do not rank the difficulty of individual shift factors. Each OOD setting uses the same checkpoint, decoding procedure, and source-calibrated thresholds.

C.2 Scoring and Operating Points

Ranking and fixed-threshold evaluation measure complementary aspects of temporal prediction. Native-track mAP averages one-to-one, confidence-ranked interval AP over tIoU thresholds 0.30:0.05:0.70. Matching is within the same video and modality; unmatched and duplicate predictions, including predictions on pristine twins, count as false positives. At the operating point, Event-IoU averages each ground-truth event's best overlap with submitted same-track intervals, assigning zero to misses. Both video score and interval confidence must strictly exceed the source threshold. Event-IoU measures coverage; mAP additionally accounts for extra predictions.

A track score is the highest post-Soft-NMS proposal confidence, $S_m(X) = \max_i \hat{c}_i^m$, or zero if no proposal exists. The branches can have different genuine-score distributions. We place their video scores on a common reference scale using $n = 256$ pristine calibration videos:

$$\begin{aligned} p_m(X) &= \frac{1 + \sum_{i=1}^{n} \mathbf{1}[S_m(X_i^{\mathrm{cal}}) \geq S_m(X)]}{n+1}, \\ S_{\mathrm{any}}(X) &= -\min\{p_a(X), p_v(X)\}. \end{aligned} \tag{4}$$

Here, X_i^{cal} denotes a pristine source-validation video and $\mathbf{1}[\cdot]$ is the indicator function. Smaller p_m indicates a more extreme score relative to genuine videos, so larger S_{any} indicates stronger evidence of manipulation in either track. At budget α, the threshold is the $(\lfloor n\alpha \rfloor + 1)$-st largest pristine calibration score, with strict exceedance; at 5% this is the thirteenth-largest score. Any-manipulation detection uses $S_{\mathrm{any}} > \tau_{\mathrm{any}}$, with its own calibrated threshold. Attribution instead uses $d_m = \mathbf{1}[S_m > \tau_m]$: $(d_a, d_v) = (0,0), (1,0), (0,1), (1,1)$ correspond to clean, audio-only, visual-only, and both-track manipulation, respectively. For an evaluation set D with ground-truth class set $\mathcal{C}_D$,

$$\mathrm{F1}_{\mathrm{attr}}(D) = \frac{1}{|\mathcal{C}_D|} \sum_{c \in \mathcal{C}_D} \mathrm{F1}_c, \tag{5}$$

where $\mathrm{F1}_c$ is the one-versus-rest F1 of class c. A prediction of a class outside $\mathcal{C}_D$ still lowers the recall of the true class. Classes without ground-truth support are excluded because their recall is undefined; including them with zero F1 would only rescale the score by $|\mathcal{C}_D|/4$. Here, $\mathcal{C}_D$ contains all four classes on ID and held-out generators, three on pooled source transfer (746 clean, 384 audio-only, and 360 visual-only videos), and two on held-out TI2V (300 clean and 300 visual-only videos). The decision-level combination leaves native event intervals and proposal scores unchanged. Recall and realized FAR are measured on each evaluation population, including distribution shifts, at this fixed source-domain operating point.

For detection, attribution, and Joint-timeline metrics, the OOD macro-average (OOD-All) weights source transfer, held-out generators, and held-out TI2V equally; for attribution, it averages per-setting scores computed over different class sets. Native OOD A-mAP averages source transfer and held-out generators, the settings with audio edits, and native OOD V-mAP averages all three settings. OOD AV-mAP averages the per-setting AV-mAP over the three settings; in each setting, AV-mAP averages A-mAP and V-mAP over the tracks with annotated events, so it equals V-mAP on held-out TI2V. Variation across training runs is reported as sample standard deviation; pair-bootstrap intervals quantify uncertainty over sampled video pairs.

C.3 Native-Track and Joint-Timeline Evaluation

Temporal overlap alone does not establish which modality was manipulated. We quantify this distinction on all 4,694 ID test videos using predictions from the three NFSR training seeds. For each model, we evaluate the complete prediction set and two subsets retaining only audio or visual

proposals, without changing their boundaries. Native-track evaluation matches predictions to ground-truth events on the same track. Joint-timeline evaluation instead merges overlapping audio and visual ground-truth intervals within each video and matches proposals without modality labels. Both use mAP@[0.30:0.05:0.70].

For joint-timeline evaluation of the complete prediction set, proposal confidences are mapped to track-specific empirical anomaly percentiles using 256 pristine source-validation videos, allowing cross-track ranking. Single-track prediction sets and native-track evaluation use the original confidences. This protocol defines the Joint-mAP reported in Table 2.

Table 11: **Modality coverage under native-track and joint-timeline evaluation.** Results are mean $\pm$ sample standard deviation over three seeds (%). A missing prediction track has zero native mAP.

Predictions	A-mAP	V-mAP	Joint-mAP
Audio + visual	92.24 ± 0.16	90.36 ± 0.28	72.33 ± 0.10
Audio only	92.24 ± 0.16	0.00	57.36 ± 0.12
Visual only	0.00	90.36 ± 0.28	64.04 ± 0.23

The visual-only subset reaches 64.04 Joint-mAP, above every visual-only baseline in Table 2, and the audio-only subset reaches 57.36, above every audio-only baseline, although each subset produces no events on the discarded track (Table 11). Native-track evaluation makes this missing coverage explicit: the absent track receives zero mAP, while performance on the retained track is unchanged. A high Joint-mAP can thus coexist with a complete modality-specific blind spot. Separate timelines preserve the distinction between locating a manipulated interval and identifying its affected track. Merging also penalizes complete predictions on both tracks, which explains why the complete model's Joint-mAP is below both native-track scores. In 727 of the 818 ID both-track videos, the audio and visual events overlap and merge into one reference interval, so the co-timed proposal from the second track counts as a duplicate false positive; without merging, the same predictions reach 91.28 mAP. Joint-mAP therefore combines boundary quality, cross-track confidence ranking, and redundancy between tracks. We use the native-track A-/V- metrics as the primary localization measures and report Joint-timeline metrics as an auxiliary view that also covers outputs without modality labels. Table 12 reports Joint-timeline metrics and native-track Event-IoU for the OOD comparison in Table 3, and Table 13 reports them for the ablation. Under Joint-mAP, NFSR also exceeds UMMAFormer in all three settings.

Table 12: **Auxiliary OOD localization metrics (%).** Joint-mAP and Joint-IoU evaluate the joint timeline, with source-transfer values averaging HDTF and TASLE with equal weight. A-IoU and V-IoU are native-track Event-IoU at the source-calibrated operating point, with the source-transfer composition of Table 3. Rows and – follow Table 3.

		Source transfer				Held-out generators				Held-out TI2V		
Method	Training Form	Joint-mAP	Joint-IoU	A-IoU	V-IoU	Joint-mAP	Joint-IoU	A-IoU	V-IoU	Joint-mAP	Joint-IoU	V-IoU
Xception	Official weights	0.86	0.42	–	0.32	0.23	0.26	–	0.19	0.25	0.12	0.12
MSLoc-PG	Official weights	26.78	19.37	–	38.53	4.62	5.78	–	9.18	0.92	2.23	2.23
ActionFormer	TASLE	37.98	20.04	–	39.87	13.89	7.54	–	10.31	15.15	7.25	6.73
Xception	AV-FABLE	1.73	0.30	–	0.47	0.41	0.35	–	0.49	0.67	0.80	0.80
DeMamba (Stage-1 window)	AV-FABLE	0.29	0.70	–	0.94	1.77	0.72	–	1.19	2.17	1.12	1.12
ActionFormer	AV-FABLE	21.38	13.32	–	26.27	33.23	24.79	–	40.03	61.57	42.93	42.93
AASIST	Official weights	0.79	0.00	0.00	–	0.27	0.24	0.42	–	0.28	0.26	–
AASIST	AV-FABLE	1.10	7.51	15.09	–	0.51	3.53	8.01	–	0.48	0.83	–
Qwen2.5-Omni-7B (ZS)	Official weights	0.89	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.86	0.00	0.00
Qwen3-Omni-30B-A3B (ZS)	Official weights	1.15	3.48	0.07	6.74	0.04	0.47	0.09	0.69	0.04	0.38	0.38
video-SALMONN 2+ 7B (ZS)	Official weights	0.00	0.00	0.00	0.00	0.00	0.04	0.11	0.07	0.00	0.03	0.03
DAV-Det	Official weights	10.53	23.42	0.00	46.60	1.12	1.14	0.14	1.05	1.18	0.92	0.81
BA-TFD+	AV-FABLE	0.85	0.99	0.65	1.32	0.55	4.86	10.15	0.26	0.07	0.61	0.46
UMMAFormer + Qwen3	AV-FABLE	61.15	50.45	81.64	22.23	34.29	15.96	26.69	15.12	22.39	4.96	12.30
Qwen3 direct timestamp SFT	AV-FABLE	15.80	26.31	52.07	1.96	6.40	10.05	24.11	1.84	0.31	0.31	1.21
AuViRe	AV-FABLE	0.21	0.00	–	–	0.39	2.39	–	–	0.07	0.75	–
ICS-Sync	AV-FABLE	1.19	0.13	–	–	1.13	1.99	–	–	0.74	0.74	–
DAV-Det + AV-FABLE	AV-FABLE	8.92	25.78	0.14	49.25	3.48	4.93	2.32	5.56	7.63	5.51	5.51
NFSR + Qwen3 (Ours)	AV-FABLE	73.70	56.22	90.86	26.42	48.55	26.03	33.40	32.92	64.92	38.83	41.21

Table 13: **Auxiliary metrics for the ablation.** Mean ± sample standard deviation over three seeds (%). OOD-All Joint-timeline metrics average the three OOD settings equally; A-IoU and V-IoU follow the native-track OOD-All definition of Table 4.

	ID		OOD-All			
Variant	Joint-mAP	Joint-IoU	Joint-mAP	Joint-IoU	A-IoU	V-IoU
w/o intermediate-layer features	73.41 ± 0.04	73.35 ± 0.83	53.22 ± 0.40	35.07 ± 2.02	56.98 ± 1.36	27.74 ± 1.68
w/o interface features	77.49 ± 1.09	81.91 ± 1.22	57.00 ± 0.26	43.46 ± 1.04	62.72 ± 1.33	36.33 ± 2.80
w/o host conditioning	72.93 ± 1.85	78.86 ± 0.58	56.80 ± 1.41	37.01 ± 2.45	58.97 ± 2.69	28.44 ± 0.66
Full NFSR	72.33 ± 0.10	81.21 ± 0.14	62.39 ± 0.21	40.36 ± 1.63	62.13 ± 0.71	33.52 ± 1.36

C.4 Readout and Baselines

The Qwen3-Omni encoders are frozen. Cached features use a 0.25-second grid: visual intermediate-layer $[N, 16, 1152]$, visual interface $[N, 2048]$, audio intermediate-layer $[N, 1280]$, and audio interface $[N, 2048]$, where N denotes the number of temporal positions. The audio intermediate layer is layer 16 (of 32) in both tested Qwen-Omni generations. The readout uses these fixed layers; layer selection is evaluated separately using training-video cross-validation (Appendix D).

Spatial pooling and gated projection. A learned linear scorer assigns a weight to each of the 16 visual cells; softmax across cells yields the weighted-sum intermediate-layer vector $\mathbf{d}_t^v$. Audio intermediate-layer vectors require no spatial pooling. With $\mathbf{q}_t^m = [\mathbf{d}_t^m; \mathbf{u}_t^m]$, the fusion is

$$\mathbf{h}_t^m = \mathrm{Dropout}(P_m(\mathrm{LN}(\mathbf{q}_t^m)) \odot \mathrm{sigmoid}(G_m(\mathbf{q}_t^m))) \,. \tag{6}$$

Here, P_m and G_m are learned affine projections to $D = 256$ channels, and LN denotes layer normalization. The gate modulates projected channels; dropout uses probability 0.1. The two branches have separate parameters.

Host statistics and modulation. For the valid temporal positions $\mathcal{T}_X$, let $M = |\mathcal{T}_X|$. Channel-wise statistics are computed as

$$\begin{aligned} \boldsymbol{\mu}_m &= \frac{1}{M} \sum_{t \in \mathcal{T}_X} \mathbf{h}_t^m, \\ \boldsymbol{\sigma}_m &= \left(\frac{1}{M} \sum_{t \in \mathcal{T}_X} (\mathbf{h}_t^m - \boldsymbol{\mu}_m)^{\odot 2} + \epsilon \mathbf{1} \right)^{1/2} . \end{aligned} \tag{7}$$

Here, powers and roots are element-wise and $\epsilon = 10^{-6}$ stabilizes the standard deviation. Padding is excluded from both statistics. A two-layer GELU MLP produces the scale and shift in Equation 1; the input video's statistics are recomputed at inference while learned parameters remain fixed.

Temporal decoding and supervision. The audio/visual pyramids have six/five levels with stride-two downsampling, masked dilated convolutions, and dynamic classification/regression heads (Zhang et al., 2022; Yang et al., 2024). A positive point lies within the event's center-sampling window (radius 1.5 level strides, clipped to the event) and satisfies the level's regression range. For level ℓ, the lower distance bound is zero at $\ell = 0$ and $2^{\ell+1}$ otherwise; the upper bound is $2^{\ell+2}$, or 10,000 at the final level, in 0.25-second grid units. The larger left/right target distance is checked against these bounds; overlapping eligible events are resolved by selecting the shortest. Regression targets are normalized by level stride. Unassigned valid points are negatives, and padded points are excluded. Classification and regression losses share a running positive-count normalizer (initial value 100, momentum 0.9). Sigmoid focal loss uses $\alpha = 0.25$ and $\gamma = 2$; regression is evaluated only at positive points. Feature fusion, host conditioning, and temporal decoding are trained jointly within each branch. Focal loss, DIoU, and Soft-NMS follow Lin et al. (2017); Zheng et al. (2020); Bodla et al. (2017). The highest-scoring valid proposals from each level enter temporal Soft-NMS with boundary voting. Retained intervals are converted to seconds and clipped to video duration. Empty or zero-length intervals are discarded. The source-validation localization score selects checkpoints before test and OOD evaluation.

Table 14: NFSR readout architecture, optimization, and decoding settings.

Setting	Value
Width / dropout	256 / 0.1
Optimizer	AdamW; learning rate 10^{-4}; weight decay 0.025
Training schedule	12 epochs; 2 warmup epochs; cosine decay; batch 8
Stabilization	Gradient clipping 1.0; EMA 0.999
Heads / assignment	Three layers; kernel 5; center-sampling radius 1.5
Losses	Sigmoid focal classification + 1D DIoU regression
Decoding	Pre-NMS 0.001; top-k 2,000; Soft-NMS $\sigma = 0.5$
Boundary voting	tIoU 0.7
Model selection	Validation mAP averaged over supported tracks
Seeds	Three training seeds
Core ablation	Same three seeds and 12-epoch budget

Baseline adaptation. Baseline inputs follow their supported modalities, with architecture-specific adaptation schedules. Xception uses two full frames per two-second cell and full-model AdamW fine-tuning (learning rate/weight decay 10^{-5}, batch 32, up to 20 epochs, validation cell-AUROC selection). The AV-FABLE-trained DeMamba Stage-1 baseline uses eight frames per two-second window, four training epochs, about 42.9k updates, and fixed proposal conversion. The MSLoc-PG row instead uses the TASLE-trained proposal generator; the released proposal-plus-refinement system is evaluated in Appendix F.4. DAV-Det adaptation trains LoRA adapters in the official audio and visual encoders for up to ten epochs with validation-AUROC selection. Both DAV-Det rows score consecutive three-second windows on each track, with four frames per visual window; each window becomes an interval proposal with its probability as confidence, and the maximum window probability gives the track score. BA-TFD+ uses official visual/audio encoders, 96^2 full frames at 2 fps, four epochs and 87,632 updates. The AV-FABLE-trained ActionFormer is trained on the complete training set for 24 epochs. UMMAFormer uses the same frozen Qwen3 features as NFSR, a 12-epoch budget, and validation-selected EMA weights. NFSR, UMMAFormer, and the TASLE-trained ActionFormer are evaluated over three training seeds; the AV-FABLE-trained ActionFormer and the other adapted baselines use one seed. Native-track scores are reported for methods producing modality-specific intervals, including timestamped language events with modality labels. Unified predictions without modality labels are scored only by joint-timeline evaluation; output types are given in the Output column of Table 2. For single-modality methods, events on the unsupported track remain misses in joint-timeline evaluation rather than being excluded.

D Forensic Evidence in Frozen Representations

We examine the representation evidence underlying NFSR through linear probes, encoder replacement, and decomposition of features across depth.

D.1 Probe Design and Layer Selection

To distinguish learned forensic discrimination from generic feature anomalies, we compare within-video normalized feature deviation, temporal feature jumps, and a supervised linear probe (Alain & Bengio, 2016). Logistic regression uses 3,000 training videos (1,500 per class), all positive cells, a 2:1 negative-to-positive cell ratio, standardized features, and L2 regularization 10^{-3}. Audio inputs concatenate intermediate-layer and interface features; the visual probe uses interface features. Evaluation measures cell-level forensic separability.

Discriminative structure beyond feature anomalies. Linear probes outperform both unsupervised scores on every track and split (Table 15). Audio AUROC reaches 90.8 on ID and 77.3 on OOD, while temporal jumps remain near chance. Supervised probes therefore recover discriminative signals along a learned feature-space direction, beyond generic deviations or changes, and these signals transfer to the OOD pool without updating the encoder.

Table 15: Cell-level forensic separability: AUROC (%). OOD denotes the OOD pool (Table 8).

Track / split	Deviation	Jump	Linear
Audio / ID	68.5	44.7	90.8
Audio / OOD	56.2	47.9	77.3
Visual / ID	62.5	49.8	80.8
Visual / OOD	64.4	53.7	68.8

Forensic accessibility across depth. Layer selection uses five-fold cross-validation grouped by training video and selects the layer with the highest held-out cell AUROC. It selects audio layer 12 in both Qwen-Omni generations: Qwen3 obtains 89.6 ± 0.7 compared with 89.1 ± 0.9 at layer 16, and Qwen2.5 obtains 86.8 ± 1.6 compared with 84.7 ± 1.5. These results locate useful audio evidence within the perception encoder and show that its accessibility varies with representation depth. For Qwen3-Omni, layers 12 and 16 differ by 0.5 AUROC, within fold variation, so NFSR retains layer 16 rather than tuning the layer on probe results (Figure 6).

The Qwen3-Omni visual sweep applies the same grouped five-fold procedure to spatially mean-pooled features from 700 training twin pairs, keeping both members of each pair in the same fold. Among 12 probed visual blocks (zero-indexed; the last is block 26), block 6 yields the highest cell AUROC: 88.44 ± 0.58 alone and 89.37 ± 0.74 when concatenated with post-merger interface features (mean $\pm$ standard deviation across folds). Later blocks fluctuate at lower values, with the final block reaching 80.55 ± 1.16 alone. Both visual variants therefore peak before the final block: greater encoder depth does not necessarily make forensic evidence more linearly accessible.

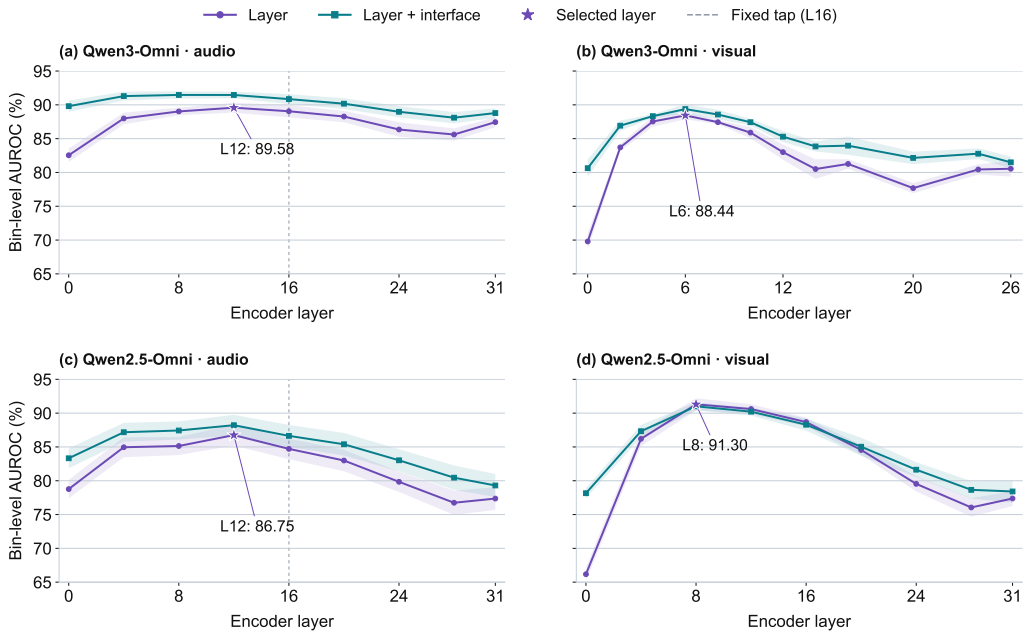


Figure 6: Layer-wise forensic separability of Qwen3-Omni (a, b) and Qwen2.5-Omni (c, d). Each panel plots cell-level probe AUROC for each layer alone and concatenated with interface features; shading shows cross-validation variation, stars mark the selected layer, and dashed lines mark the fixed audio layer (L16). Visual features are spatially pooled; audio and visual depth indices are separate.

D.2 Encoder Replacement

We apply temporal readouts to frozen encoders with different pretraining objectives to assess how the available representation affects localization and transfer. Table 16 reports native-track mAP so that audio and visual behavior can be examined separately. Both Qwen-Omni generations support

Table 16: Frozen-encoder comparison: native-track mAP (%). HDTF reports audio localization on the held-out talking-head domain; generator and TI2V columns report their respective modality-specific evaluations.

Encoder	ID A	ID V	HDTF A	Gen. A	Gen. V	TI2V V
Whisper-large-v3 (Radford et al., 2023)	77.3	–	86.8	31.4	–	–
BEATs (Chen et al., 2023) + CLIP (Radford et al., 2021)	93.9	83.4	85.9	56.1	49.2	49.6
Qwen2.5-Omni-7B	87.5	93.7	94.6	37.2	58.8	72.3
Qwen3-Omni-30B	92.2	90.4	98.4	48.9	49.3	68.1

localization on held-out hosts and editing configurations. Qwen3 leads on HDTF audio, Qwen2.5 on the visual evaluations, and BEATs+CLIP on ID and generator-shifted audio. Thus, forensic readout extends across these encoders, with complementary strengths by modality and domain. Qwen2.5/Qwen3 use two/three seeds; their generator-shift audio/visual mAP standard deviations are 1.5/3.3 and 1.7/0.7, respectively.

D.3 Complementary Evidence across Representation Stages

To characterize the evidence shared between representation stages, we decompose audio intermediate-layer features into a component predicted from the interface and a residual. A ridge map $E : \mathbb{R}^{2048} \to \mathbb{R}^{1280}$ is fitted from interface features $\mathbf{a}^s$ to intermediate-layer features $\mathbf{a}^d$ using 636,000 pristine training cells. The regularizer is $10^{-3}\,\mathrm{tr}(\Sigma_{ss})/2048$. It separates a linearly reconstructed component from a residual:

$$\mathbf{e}_t = E(\mathbf{a}_t^s), \qquad \mathbf{r}_t = \mathbf{a}_t^d - E(\mathbf{a}_t^s). \tag{8}$$

Here, $\mathbf{e}_t$ is the component reconstructed from interface features and $\mathbf{r}_t$ is the remaining intermediate-layer feature vector. At layer 16, reconstruction on held-out pristine cells gives $R^2 = 0.71$. Linear probes evaluate the two components and their concatenations with the interface. Temporal decoding is also evaluated with the residual in place of the raw intermediate-layer features under a 12-epoch training budget.

Table 17: Audio representation decomposition. Top: cell-level probe AUROC (%). Bottom: temporal mAP (%) under a 12-epoch budget. Intermediate + interface reports three-seed mean and standard deviation; alternatives use one seed. EPIC and ANet-gen denote held-out-generator edits on EPIC-KITCHENS and ActivityNet hosts; OOD pool is defined in Table 8.

Probe input	ID	HDTF	XTTS-v2	EPIC	ANet-gen
Interface $\mathbf{a}^s$	86.1	84.6	84.2	84.8	68.7
Intermediate $\mathbf{a}^d$	89.4	89.1	93.2	83.4	71.3
Residual $\mathbf{r}$	84.1	84.8	88.0	71.5	61.0
Reconstructed $E(\mathbf{a}^s)$	86.0	85.3	85.3	84.6	68.8
$[\mathbf{r}, \mathbf{a}^s]$	90.6	92.3	93.0	84.4	72.1
$[\mathbf{a}^d, \mathbf{a}^s]$	90.7	92.4	93.2	85.1	71.9
Residual norm $\|\mathbf{r}\|_2^2$	45.7	48.1	51.5	61.5	47.8
Decoder input	**ID**	**HDTF**	**XTTS-v2**	**Omni2Sound**	**OOD pool**
Intermediate + interface	92.2 ± 0.2	98.4 ± 0.6	92.3 ± 0.6	41.4 ± 1.1	85.6 ± 1.0
Intermediate only	91.5	98.5	91.2	37.8	84.4
Residual + interface	88.0	93.2	82.3	28.6	76.6

Predictive information spans both components. The residual remains discriminative on HDTF and XTTS-v2, but its norm is a weak authenticity score (Table 17). The relevant signal is therefore structured within the residual rather than captured by its magnitude. The reconstructed component is also predictive, and the original intermediate-layer features outperform the residual alone on every listed domain. Concatenating intermediate-layer and interface features further improves probe AUROC on ID, HDTF, and ANet-gen, consistent with complementary evidence at the two stages.

Feature accessibility and temporal decoding are distinct. For a fixed E, $[\mathbf{r}, \mathbf{a}^s]$ and $[\mathbf{a}^d, \mathbf{a}^s]$ are invertible linear reparameterizations with similar probe scores. Nevertheless, residual inputs

reduce the temporal decoder's OOD-pool mAP from 85.6 to 76.6 under the tested budget. A linearly information-preserving transformation need not be equally effective for temporal decoding, supporting the use of the original features rather than subtraction of their shared component.

E Sparse Evidence in Long Videos

Sparse manipulation requires preserving local evidence while controlling false alarms over long authentic context. Mean aggregation and local triggering exhibit complementary failure modes in this setting. Consider a video of T temporal cells with a manipulated event occupying $L \ll T$ cells. Suppose manipulated cells have a local evidence mean shift δ, and the whole-video mean has variance σ^2/T_{eff}, where T_{eff} summarizes temporal dependence. The standardized separation of the mean scores is

$$z(T) = \frac{\delta L \sqrt{T_{\text{eff}}}}{\sigma T}. \tag{9}$$

For independent cells, $T_{\text{eff}} = T$, giving $z(T) = \delta L/(\sigma\sqrt{T})$. With fixed event length, adding authentic context reduces this mean-score separation. Under an additional equal-variance Gaussian score model, $\mathrm{AUROC} = \Phi(z/\sqrt{2})$, where Φ is the standard normal cumulative distribution function.

At the opposite extreme, triggering on any local alarm creates more opportunities for false positives. For N independent decisions with local false-positive probability q,

$$\Pr(\text{one or more false alarms}) = 1 - (1-q)^N \simeq Nq, \qquad Nq \ll 1. \tag{10}$$

Here, N counts local decisions, not necessarily independent video frames; temporal correlations change the effective number of alarm opportunities. Mean aggregation can dilute a fixed-length event, whereas unrestricted local triggering can accumulate false alarms as context grows. These stylized cases motivate retaining evidence at event scale while evaluating false alarms at video scale. NFSR follows this design through temporal proposals and source-validation calibration of the resulting video scores.

F Generalization and Sensitivity Analyses

We analyze training variability, supervision coverage, external transfer, and robustness to degraded inputs.

F.1 Seed Stability

We repeat the shared-feature comparison across three seeds to assess whether the readout advantage persists under training variation. NFSR's ID advantage over UMMAFormer (6.54 AV-mAP points,

Table 18: Three-seed stability: mean ± sample standard deviation. ID and OOD localization are evaluated on the native audio and visual tracks. OOD AUROC averages the three OOD settings equally; OOD AV-mAP, A-mAP, and V-mAP follow the OOD-All definitions of Table 4.

Method	ID AUC	ID AV-mAP	ID A-mAP	ID V-mAP	ID A-IoU	ID V-IoU	OOD AUC	OOD AV-mAP	OOD A-mAP	OOD V-mAP
ActionFormer (TASLE)	67.94±0.15	5.06±2.59	–	10.13±5.18	–	5.56±2.25	72.70±0.05	21.06±8.31	–	37.07±14.00
UMMAFormer + Qwen3	89.18±0.29	84.76±0.19	89.90±0.22	79.62±0.16	79.77±1.90	65.25±0.20	70.33±0.80	43.50±0.64	69.94±1.50	28.36±1.17
NFSR + Qwen3	96.04±0.10	91.30±0.13	92.24±0.16	90.36±0.28	84.72±0.55	76.90±0.78	84.61±0.12	62.45±0.08	73.66±0.79	53.10±0.56

with 2.34 A-mAP and 10.74 V-mAP points) exceeds the standard deviations of these mAP values, which are at most 0.28 (Table 18). The OOD localization advantage (18.95 AV-mAP points, with 3.72 A-mAP and 24.74 V-mAP points) likewise exceeds the observed standard deviations of at most 1.50, supporting the stability of the shared-feature comparison. The TASLE-trained ActionFormer has stable detection AUROC but localization that varies substantially across seeds, particularly under distribution shift.

F.2 Value of AV-FABLE Supervision

We compare visual-only NFSR readouts trained on TASLE and on AV-FABLE to examine the effect of the training corpus on visual-event localization. Both readouts keep only the visual branch and

are trained with visual supervision alone, matching TASLE, which contains only visual edits; their values therefore differ from those of the audio-visual NFSR in the main tables. Both are evaluated on AV-FABLE and TASLE; DAV-Det provides a separate test of adapting a video-authenticity detector to AV-FABLE. Corpus coverage strongly affects transfer (Table 19). The TASLE-trained visual readout

Table 19: Training-corpus comparison (%). All rows report visual-track detection AUROC, which separates videos with visual events from pristine videos, and V-mAP. Models are trained separately on the listed corpora; both NFSR rows are visual-only readouts, not the audio-visual model of Table 2. OOD denotes the OOD pool. TASLE cells are AUROC / visual mAP.

Model / training corpus	ID AUC	ID V-mAP	OOD AUC	OOD V-mAP	TASLE
NFSR (visual-only) / TASLE	73.2	25.2	83.0	38.8	98.30 / 93.71
NFSR (visual-only) / AV-FABLE	96.9	91.0	87.5	60.3	84.4 / 59.0
DAV-Det / MVAD	52.77	0.21	64.98	0.84	78.87 / 21.00
DAV-Det / AV-FABLE	86.33	2.21	68.73	2.98	70.46 / 17.76

performs well on TASLE but reaches only 25.2 visual mAP on AV-FABLE ID events. AV-FABLE supervision raises this to 91.0 and OOD-pool mAP from 38.8 to 60.3, while transferring to TASLE at 59.0 mAP without target-dataset training. AV-FABLE supervision thus transfers to an external benchmark, although TASLE-trained models remain stronger on their own benchmark. Adapting DAV-Det mainly improves ID detection (52.77 to 86.33 AUROC), with a smaller OOD gain and lower TASLE scores, while its visual mAP remains below 3 on AV-FABLE. The localization benefit of AV-FABLE supervision therefore also depends on the prediction model.

F.3 Training-Set Scale

We vary the fraction of AV-FABLE training data while retaining the same epoch schedule. This measures the practical benefit of a larger training resource under a fixed-epoch schedule. Both tracks

Table 20: Training-fraction sensitivity: native-track mAP (%). Fractions use the same epoch count, so sample count and optimizer updates vary together. OOD denotes the OOD pool. Entries are single-run results.

Training fraction	ID A	OOD A	ID V	OOD V
5%	55.7	16.8	5.6	5.1
10%	70.5	50.4	56.1	17.4
25%	83.6	74.7	80.0	39.3
50%	88.3	79.8	85.7	51.2
100%	92.1	84.6	90.7	55.2

improve throughout the sweep (Table 20), with particularly weak visual localization at the smallest fraction. Increasing training from 50% to the full set still raises pooled OOD audio/visual mAP by 4.8/4.0 points. Additional data and the associated optimizer updates therefore remain useful beyond the initial gains obtained with a small subset.

Generator coverage. In three generator-holdout folds, the held-out generator's route remains covered by other generators in training, yet localization mAP drops by 16.6, 17.0, and 32.8 points. Coverage of an editing route therefore does not by itself ensure transfer to the held-out generators tested here.

F.4 External Comparison with Official MSLoc

To test transfer under another benchmark's task definition, we evaluate MSLoc-PG and its released refinement checkpoint (Feng et al., 2026) alongside AV-FABLE-trained NFSR on the same 1,442 TASLE/TVSum videos with the official evaluator. $F1_{\mathrm{Det}}$ denotes its `Loc_F1` on a 0.01-second grid; $F1_{\mathrm{Loc}}$ averages interval F1 at tIoU 0.1/0.3/0.5/0.7. These are distinct from AV-FABLE video-level detection and mAP. MSLoc-PG leads on its TASLE source test, whereas NFSR leads both metrics on the pure Kling/Jimeng subset and localization F1 on TVSum (Table 21). NFSR thus transfers without

Table 21: Cross-dataset evaluation under the official TASLE protocol: $F1_{\mathrm{Det}}/F1_{\mathrm{Loc}}$ (%). Counts are manipulated/pristine. Commercial subsets overlap. Evaluation uses the full TASLE official test; the main source-transfer evaluation uses a separate source-disjoint subset.

Subset	NFSR	MSLoc-PG	PG + refine
TASLE official (631/662)	51.01 / 50.80	67.24 / 63.62	36.14 / 36.93
Commercial pool (208/662)	46.33 / 47.40	58.14 / 49.28	43.79 / 47.09
Contains Kling/Jimeng (94/662)	44.10 / 46.61	50.55 / 36.46	35.77 / 38.92
Pure Kling/Jimeng (78/662)	48.02 / 54.03	42.26 / 30.85	35.23 / 39.76
TVSum (72/77)	38.42 / 43.75	48.01 / 38.96	36.24 / 39.22

TASLE training, while source-trained MSLoc-PG remains stronger on the full test. The released refinement checkpoint does not improve aggregate TASLE scores in this evaluation.

F.5 Robustness to Input Degradation

We test whether degradation of one input track disrupts localization on that track or propagates to the other. Thresholds remain fixed by clean source-domain calibration. Table 22 reports the most severe tested condition from each corruption family on 512 videos (256 pairs), and missing-modality results on the 2,004-video OOD pool. Each block includes its own clean reference. Single-track

Table 22: Robustness to media corruption and missing modalities (%). Recall and FAR use fixed source-domain thresholds. Each evaluation population has its own unperturbed reference.

Condition	AUROC	Recall	FAR	Attr. F1	A-mAP	V-mAP
Media corruptions: 512-video subset						
Clean	88.6	58.2	2.7	61.0	85.6	53.1
H.264 CRF 35	85.1	46.9	2.0	54.0	85.6	26.0
360p	80.7	40.6	1.2	47.1	85.6	10.9
AAC 48 kb/s	83.4	39.8	2.3	56.0	65.5	53.1
Audio noise 0.2	76.2	30.1	4.3	42.8	29.1	53.1
Gain -12 dB	87.3	53.1	2.3	60.7	81.9	53.1
A/V offset 500 ms	88.5	56.6	3.1	61.8	82.3	53.1
Missing modalities: 2,004-video OOD pool						
Audio + visual	88.9	57.6	3.0	58.5	84.6	55.2
Audio missing	67.0	18.5	1.8	32.6	0.4	55.2
Visual missing	81.2	39.6	1.3	40.0	84.6	3.9

degradation affects the corresponding branch: 360p reduces visual mAP from 53.1 to 10.9 without changing audio mAP; audio noise similarly preserves visual localization. Removing a modality also preserves the other track, consistent with separate evidence paths. Gain changes and A/V offsets are less damaging in the tested ranges. Reduced recall can coexist with low FAR, so fewer false alarms alone do not imply that manipulated events remain detectable.

F.6 Cross-Track False Alarms

Audio-only speech replacement provides a paired-input control for modality specificity. In the training set, speech replacements co-occur with visual edits; the control instead changes the soundtrack while preserving the visual content. A visual prediction should therefore respond in the same way to each edited video and its pristine twin.

We compare NFSR's separated readouts with a cross-modal readout whose audio and visual branches exchange information through gated cross-modal attention at each shared pyramid level; each model uses its own source-calibrated 5% threshold. The separated visual readout fires on 2.9% of both audio-edited videos and pristine twins for all three TTS systems. The cross-modal readout's visual head instead fires on 4.6% of pristine twins but 11.8–15.1% of audio-only edits. This sensitivity to soundtrack changes indicates cross-modal coupling in the cross-modal readout's visual decisions. Separate branches preserve visual specificity when manipulation labels co-occur during training.

G Language Verification and Temporal Decisions

We examine whether language judgments improve temporal prediction through direct outputs, fixed-candidate verification, or score fusion. These analyses use an earlier training release that contains event-sound and ambience manipulations but no speech replacement, so HDTF here combines source and manipulation-type shifts. Event prediction, verification, and rescoring use event F1, classification AP, and candidate-set localization AP, respectively.

G.1 Zero-Shot and Direct-Transfer Results

We first compare direct language outputs with a trained temporal readout on frozen BEATs+CLIP features. This tests whether transferring perceptual features to temporal prediction is useful beyond the Qwen-based readouts in the main comparison. Direct language outputs yield low event F1 across

Table 23: Zero-shot (ZS) and direct-transfer (DT) evaluation (earlier release). Triples are detection Macro-F1 / visual Event-F1@0.5 / audio Event-F1@0.5 (%). HDTF combines source and manipulation-type shifts in this protocol.

Method	Source	Generator	TI2V	OOD pool	Attr. F1
Qwen2.5-Omni (ZS)	46.3 / 0.0 / 3.4	45.1 / 4.4 / 3.8	46.3 / 3.7 / 2.2	46.0 / 4.0 / 2.8	8.0
Qwen3-Omni (ZS)	35.6 / 0.0 / 0.5	42.9 / 0.6 / 2.3	42.4 / 2.4 / 0.4	42.6 / 1.7 / 1.0	18.6
video-SALMONN 2+ (ZS)	15.5 / 0.0 / 0.0	1.2 / 0.0 / 0.0	10.5 / 0.0 / 0.0	7.9 / 0.0 / 0.0	5.6
MSLoc-PG (DT)	35.1 / 0.0 / –	52.4 / 11.6 / –	43.4 / 4.6 / –	46.6 / 7.6 / –	–
BEATs+CLIP / joint readout	38.4 / 22.2 / 3.5	70.2 / 48.6 / 53.8	54.8 / 51.0 / 3.4	60.2 / 49.9 / 20.5	41.9

shifts (Table 23). The BEATs+CLIP readout recovers more events, but audio transfer is weaker on unseen HDTF speech replacements than generator-shifted event sounds. In this protocol, transfer varies with both the prediction path and training manipulation coverage. The MSLoc row evaluates proposals only; Table 21 evaluates the complete released pipeline.

G.2 Fixed-Candidate Verification

Holding candidate intervals fixed separates the ability to judge an event from the ability to propose its boundaries. We compare zero-shot verification, candidate-slot SFT, and decision-stage GRPO (Shao et al., 2024) on 130 development and 228 OOD-probe examples. Candidate authenticity AP measures verification quality; integration into the temporal pipeline tests whether better verification improves the final localized events. SFT raises mean OOD candidate AP from 39.1 to 53.4; GRPO adds a

Table 24: Candidate verification and complete-system evaluation (%). Top: slot AP is audio / visual / mean, and video metrics are AUROC / AP. Bottom: each track reports mAP / Event-F1@0.5 / positive-video mIoU / pristine FAR. Mean mAP averages the two native-track mAP values.

Verifier	Dev slot AP	Dev video	OOD slot AP	OOD video
Zero-shot	56.9 / 24.3 / 40.6	71.7 / 75.2	24.7 / 53.5 / 39.1	64.1 / 63.1
Slot SFT	92.9 / 76.6 / 84.7	89.0 / 89.8	44.5 / 62.3 / 53.4	73.3 / 71.9
Decision GRPO	92.8 / 81.5 / 87.1	91.0 / 91.3	43.8 / 63.4 / 53.6	72.0 / 71.6

System	Visual metrics	Audio metrics	Mean mAP	Attr. F1 / Acc.
Verifier off	69.08 / 65.23 / 54.22 / 4.13	78.15 / 73.79 / 67.94 / 7.87	73.61	66.56 / 77.79
+ Zero-shot	69.08 / 65.23 / 54.22 / 4.13	78.15 / 73.52 / 66.84 / 7.68	73.61	66.41 / 77.59
+ Slot SFT	69.10 / 64.86 / 53.79 / 4.33	78.10 / 72.53 / 64.50 / 6.50	73.60	66.23 / 77.18
+ Decision GRPO	69.08 / 65.23 / 54.22 / 4.13	78.17 / 72.44 / 65.36 / 7.48	73.62	66.71 / 77.48

development-set gain but little OOD improvement (Table 24). Complete-system mean mAP remains within 73.60–73.62 despite improved candidate discrimination. Better candidate verification therefore does not translate into better localization: the verifier can only rescore candidates whose boundaries and coverage are fixed by the temporal readout.

G.3 Language Rescoring and Score Fusion

We next retain the candidate intervals and compare their numerical readout scores with language-model scores, within-video standardization, and weighted fusion. This tests whether language judgments provide useful ranking information even when they do not change event boundaries. Numerical readout scores outperform language scores on every slice, and equal-weight fusion lowers

Table 25: Candidate-set AP at tIoU 0.5 (earlier release, %). Readout: numerical readout score; Readout-z: within-video standardized readout score; LM: Qwen3-Omni language-model score; Equal: equal-weight fusion; z+LM: standardized fusion; 0.8-Readout: fusion with readout weight 0.8.

Slice	Track	Readout	Readout-z	LM	Equal	z+LM	0.8-Readout
ID	A	91.4	80.9	35.6	88.6	89.1	90.8
ID	V	92.9	71.5	50.0	88.6	87.7	92.0
OOD pool	A	42.8	46.3	15.5	37.2	51.4	41.6
OOD pool	V	63.6	41.1	31.3	56.6	56.5	63.2
HDTF	A	54.4	61.2	19.0	40.2	59.6	47.5
ANet OOD	A	40.6	17.9	10.8	35.8	34.7	39.7
ANet OOD	V	65.3	48.2	32.6	57.3	58.1	64.5
Omni2Sound	A	43.0	20.0	16.7	38.3	40.5	42.2
Wan-FLF2V	V	47.9	15.9	30.3	45.4	40.4	48.2
OmniWeaving	V	32.4	13.0	18.7	28.5	24.1	32.6
TI2V	V	65.1	36.1	35.0	58.1	54.3	64.5

AP throughout (Table 25). Standardized fusion improves pooled OOD audio AP from 42.8 to 51.4 but reduces both ID tracks and OOD visual AP. Language judgments supplement particular domains without providing a consistently better ranking rule.

H Explanation Baselines

We evaluate zero-shot and LoRA-tuned Qwen3-Omni explanation baselines conditioned on fixed events. We measure output validity, temporal agreement with reference intervals, operation/target identification, and abstention. These results show that event-conditioned explanation remains challenging: output validity and content correctness diverge, semantic identification remains low, and the tuned model abstains on most manipulated inputs.

H.1 Training and Evaluation Setup

For a conditioning event (m_i, b_i, e_i, c_i) in a video of duration T, the Qwen3-Omni Thinker receives its modality, interval, confidence, and the audio-visual crop

$$W_i = [\max(0, b_i - \delta),\ \min(T, e_i + \delta)], \qquad \delta = 2\,\mathrm{s}. \tag{11}$$

Here, b_i, e_i are absolute timestamps supplied by the selected candidate or the annotated event, depending on the evaluation condition. Clipping keeps the context window within the video, while both modalities remain available for contextual explanation. A structured response carries event identity, track, absolute start/end times, operation, target type, inside/context evidence, judgment, and a support verdict. Identity, modality, and interval are copied from the conditioning event; complete abstention requires the designated none-valued semantic fields and a false support verdict. Manipulated-event examples supply evidence records; their corresponding pristine windows supply structured abstention targets. LoRA training updates the language adapter independently of NFSR's temporal readouts. Generated text leaves predicted intervals, confidence scores, detection decisions, and modality attribution unchanged.

SFT uses 4,000 examples (2,000 event–pristine pairs from 1,921 hosts), 500 steps, and LoRA rank eight/alpha 32 on all linear layers (Hu et al., 2022). Supervision is drawn from the 13,628 training events with candidate evidence records and their corresponding pristine windows. Training and evaluation hosts are disjoint.

Evaluation uses 64 ID pairs and 64 OOD pairs with duration at most 240 seconds, stratified by track, source, and generator. Zero-shot and SFT-predicted modes receive identical NFSR candidates. GT conditioning supplies annotated intervals to evaluate explanation generation with ground-truth temporal input. Strict schema scoring checks the full raw JSON. Events match on track and tIoU at least 0.5; operation/target macro-F1 retains parse failures in the fixed matched-event denominator. Temporal agreement requires both endpoints within one second of ground truth. Clean abstention measures complete abstention on pristine inputs, whereas clean support counts erroneous support verdicts; malformed or incomplete outputs can satisfy neither condition.

Table 26: Event-conditioned explanation. Scores are percentages except matched-event counts. Op./target denote operation/target-type Macro-F1; time measures boundary agreement with the reference interval. Clean support denotes erroneous support verdicts on pristine videos. SFT-GT conditions on annotated intervals, separating explanation evaluation from predicted localization.

Mode	Schema	Matched	Op./target	Time	Clean abst.	Clean support	Fake abst.
Zero / ID	82.81	62 / 64	6.00 / 11.04	72.58	20.31	1.56	18.75
Zero / OOD	91.41	49 / 64	3.10 / 19.85	87.76	17.19	3.13	17.19
SFT / ID	99.22	62 / 64	16.38 / 20.91	79.03	87.50	10.94	67.19
SFT / OOD	100.00	49 / 64	4.44 / 15.11	93.88	93.75	6.25	76.56
SFT-GT / ID	100.00	64 / 64	13.91 / 13.47	100.00	89.06	10.94	76.56
SFT-GT / OOD	100.00	64 / 64	1.28 / 13.64	100.00	92.19	7.81	79.69

H.2 Structured Output and Semantic Identification

SFT produces valid raw structured outputs on 99.22% of ID and 100.00% of OOD inputs (Table 26). On ID events, operation/target F1 increases from 6.00/11.04 to 16.38/20.91 relative to zero-shot generation. The corresponding OOD results are mixed: operation F1 increases slightly while target F1 decreases. Format compliance and event-label accuracy therefore respond differently to supervision.

Abstention and event support. SFT raises pristine abstention to 87.50% on ID and 93.75% on OOD, but also abstains on 67.19% and 76.56% of manipulated inputs. It thus learns strong abstention without reliably distinguishing unsupported candidates from manipulated events.

Conditioning on annotated intervals. With annotated events, SFT-GT achieves complete matching and 100% boundary agreement by construction, while operation/target F1 remains low and abstention frequent. Accurate temporal conditioning alone is therefore insufficient for semantic identification.

Prompt and formatting sensitivity. Zero-shot schema failures stem from outer Markdown code fences; removing them makes every zero-shot response schema-compliant without changing its content. Clarifying numeric types and absolute timestamps in the SFT prompt yields no reliable operation-F1 change ($+3.44$ on ID, pair-bootstrap interval $[-0.84, 10.67]$; 0.00 on OOD, $[-4.52, 4.43]$).

H.3 Media Dependence of Support Decisions

To test whether support decisions depend on the supplied media, we compare three input conditions using the same SFT-GT model and the 64 ID and 64 OOD twin pairs. Prompts, event identities, modalities, and annotated intervals are held fixed. The original condition supplies the event-centered crop. The shifted condition supplies an event-free window from the same video, using matching windows in each twin; shorter windows are padded with repeated terminal frames and silence to preserve crop duration. The target-track-masked condition retains the original crop but replaces the queried modality with silence or black frames, leaving the other modality unchanged.

Replacing the event crop with an event-free window reduces support by 14.06 percentage points on ID and 15.63 on OOD (Table 27). The respective 95% paired-bootstrap intervals are $[-28.13, -1.56]$ and $[-26.56, -6.25]$, using 1,000 resamples of twin pairs shared across input conditions. Masking the queried track eliminates support on both splits. Thus, the supplied event prompt alone does not determine the support verdict, which is sensitive to event content and the availability of the queried

Table 27: **Support decisions under media interventions.** Rates (%) are measured separately on the manipulated and pristine members of each twin pair, with 64 pairs per split. Support and complete abstention require valid structured outputs and unchanged conditioning fields. Original crops use the SFT-GT setting of Table 26.

	Manipulated-twin support		Pristine-twin abstention	
Media condition	ID	OOD	ID	OOD
Original crop	23.44	20.31	89.06	92.19
Event-free shifted crop	9.38	4.69	92.19	98.44
Target track masked	0.00	0.00	100.00	100.00

modality. Sensitivity to the media does not establish that the verdicts or evidence statements are correct; the operation/target scores in Table 26 remain low.

I Qualitative Examples

Figure 7 illustrates manipulated events on both tracks: object replacement and insertion on the visual track, and sound-event and speech replacement on the audio track. Rows (c) and (d) show co-timed visual and audio events in the same both-track video.

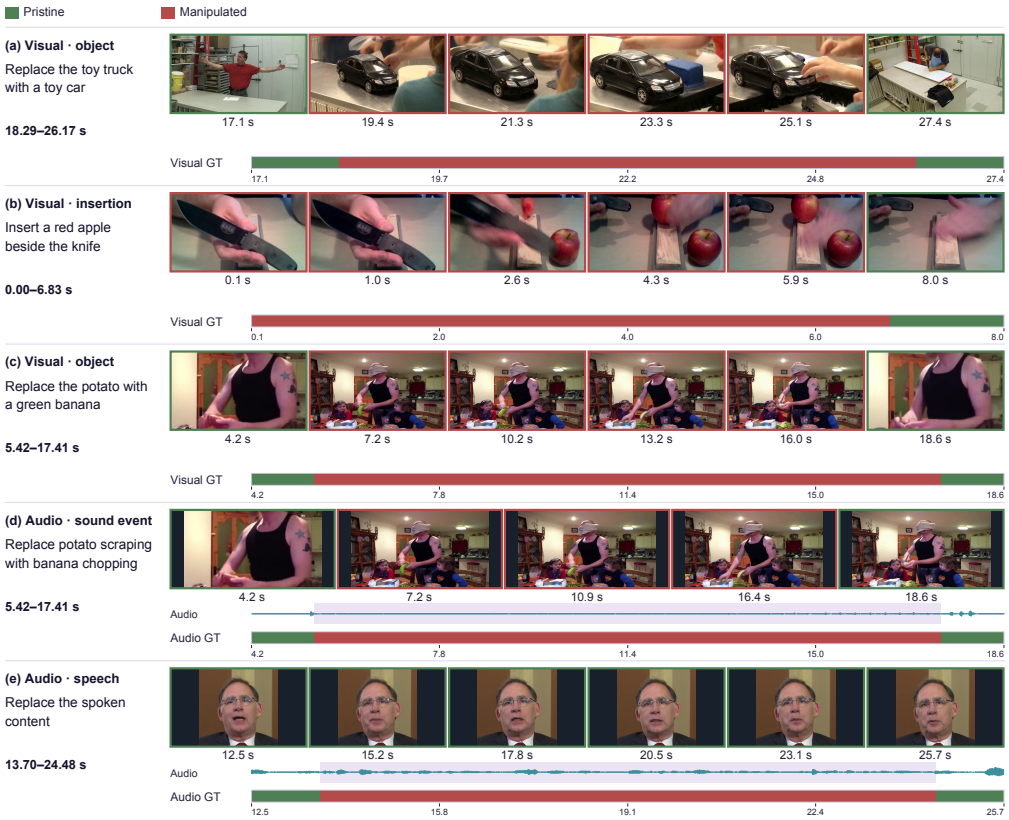


Figure 7: Manipulation examples in AV-FABLE. Each row gives a short description of the edit, the annotated interval, frames sampled across the event boundaries, and the local ground-truth (GT) timeline of the manipulated track; audio rows add the manipulated waveform with the event shaded. Green and red denote pristine and manipulated content, and frame borders follow the visual track. Rows (c) and (d) are co-timed events in one both-track video.

Figures 8 and 9 show NFSR predictions at the fixed source-calibrated track thresholds. Figure 8 contains a localized visual-only object replacement, a localized audio-only speech replacement, a

missed visual object insertion, and a false alarm on a pristine video. Figure 9 shows a both-track video with four asynchronous events, two on each track. NFSR matches both audio events and the second visual event (tIoU 0.753–0.973) but misses the first visual event.

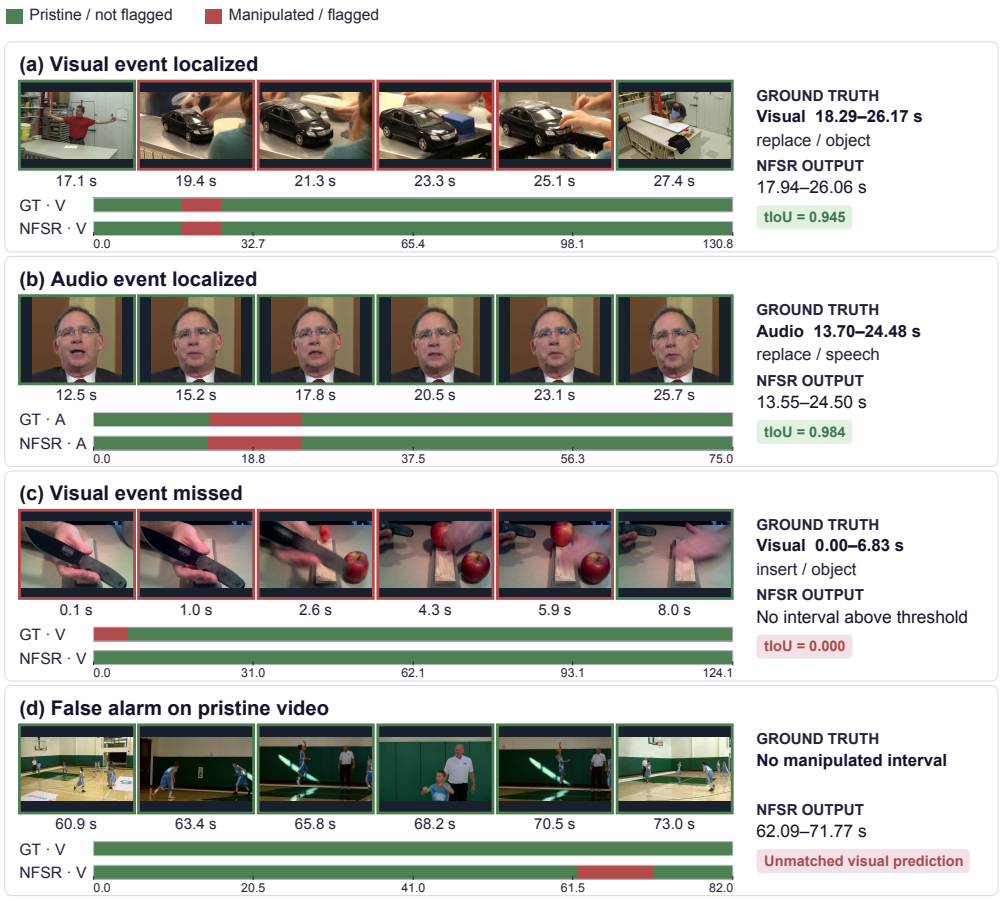


Figure 8: Localization successes and errors. (a) A visual-only object replacement and (b) an audio-only speech replacement are localized; (c) a visual object insertion is missed; (d) a pristine video yields a visual false alarm. Each row shows sampled frames, full-video GT and NFSR timelines of the displayed track, the annotated and predicted intervals, and the best tIoU. NFSR intervals are those above the fixed track threshold; frame borders follow the visual-track GT.

Figure 10 shows three ID examples from the evaluation in Table 26. Zero-shot and SFT-predicted modes receive the same top-ranked NFSR candidate, taken before the video-level decision threshold is applied; candidate tIoU is 0.857–0.991. Zero-shot generation withholds support in all three examples, whereas SFT supports each event with inside, context, and judgment fields. These supported cases are selected for illustration; in aggregate, SFT often abstains on manipulated inputs (Table 26). The displayed fields are model outputs rather than verified evidence; they illustrate the distinction between locating an event and describing its forensic evidence.

Data access. Dataset access will follow the licenses of the underlying source collections. Where direct media redistribution is restricted, access will be supported through permissible annotations, source identifiers, and temporal coordinates.

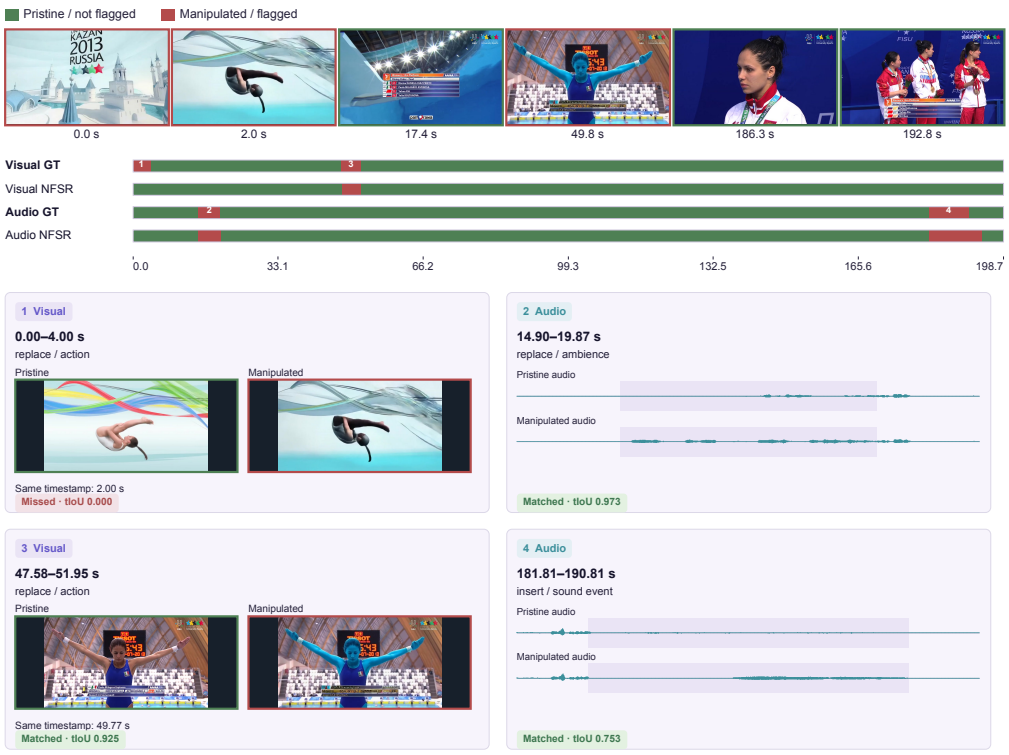


Figure 9: Four asynchronous events in one both-track video. Top: sampled frames and full-video GT and NFSR timelines for each track, with numbered GT events. Bottom: each event's interval and operation/target, pristine and manipulated frames at the same timestamp (visual) or pristine and manipulated waveforms over the event window (audio), and the best tIoU. NFSR matches events 2–4 (tIoU ≥ 0.5) and misses event 1.

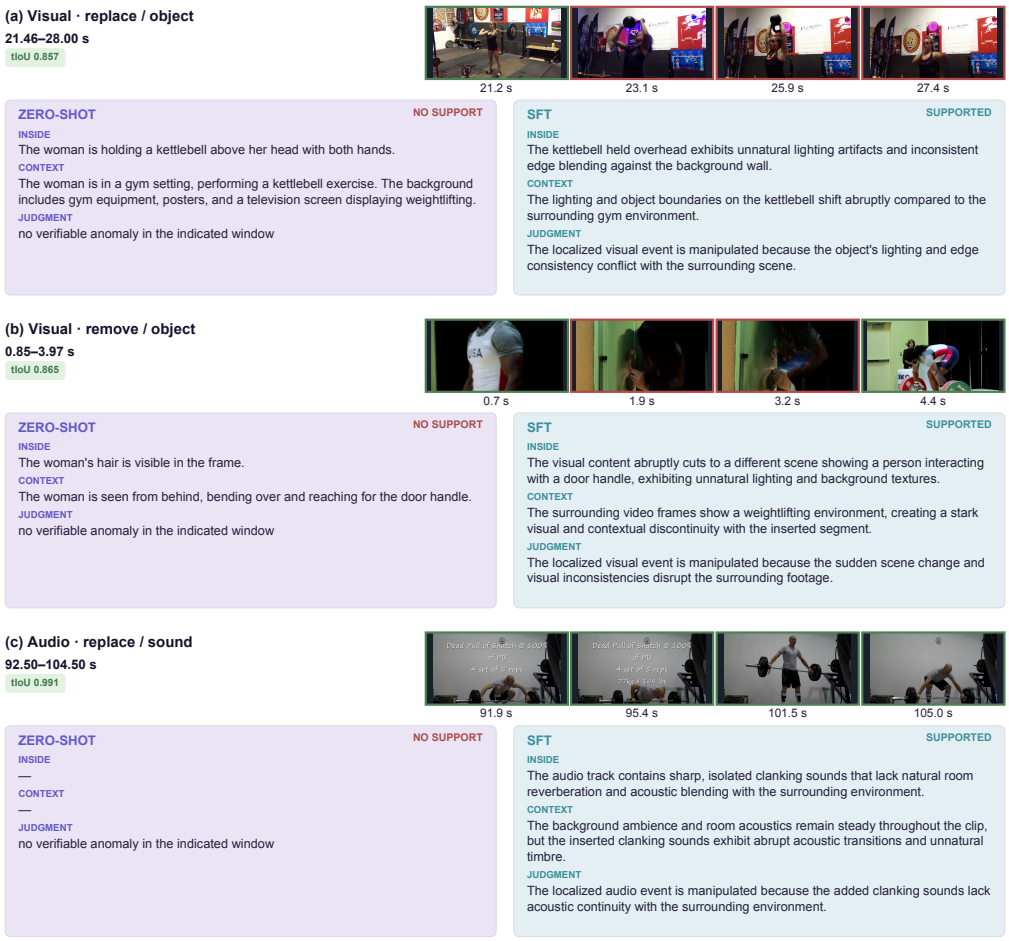


Figure 10: Explanation outputs on fixed NFSR candidates. Each example gives the annotated track, operation, target, and interval, the tIoU of the candidate supplied to both modes, and sampled frames. Zero-shot and SFT panels show the support verdict and the generated inside, context, and judgment fields; dashes mark empty fields.