EVIPATROL: EVIDENCE-GROUNDED SPLIT-MODAL MLLM FOR AUDIO-VISUAL DEEPFAKE LOCALIZATION AND DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-visual deepfake localization requires identifying the manipulated modality, precise temporal boundaries, and supporting evidence. However, existing methods often retain localization-related cues implicitly within latent representations or prediction scores, limiting their interpretability and downstream reasoning. Meanwhile, multimodal large language models (MLLMs) struggle to capture weak and localized boundary signals from raw audio-visual inputs due to their emphasis on semantic understanding. We propose **EviPatrol**, an explicit boundary evidence modeling framework for fine-grained audio-visual deepfake localization and diagnosis. EviPatrol transforms implicit localization cues into structured, time-indexed evidence representations through a Boundary Evidence Interface before MLLM reasoning. To improve evidence reliability, we introduce BLA-AMR and Safe-AMR to enhance modality-aware evidence learning and reduce cross-modal shortcuts. The resulting evidence is integrated with audio-visual content through a split-modal MLLM for independent audio and visual localization with evidence-grounded diagnosis. Extensive experiments on LAV-DF and AV-Deepfake1M demonstrate that EviPatrol achieves superior temporal localization and cross-dataset generalization. On LAV-DF, EviPatrol improves AP@0.95 by **51.94** percentage points over the previous best method, validating the effectiveness of explicit boundary evidence modeling under strict localization criteria.

1 INTRODUCTION

Recent advances in generative AI have enabled highly realistic audio-visual manipulations, where only a few spoken words or short facial motion segments can be modified while preserving the remaining content as authentic (Heidari et al., 2024; Xie et al., 2026). Such localized manipulations pose significant challenges for forensic analysis, as short forged intervals may compromise the authenticity of an otherwise genuine recording. Therefore, deepfake analysis requires not only identifying whether a video is manipulated, but also determining the manipulated modality and accurately localizing the temporal intervals. Recent benchmarks, including LAV-DF (Cai et al., 2022) and AV-Deepfake1M (Cai et al., 2024), have established fine-grained audio-visual deepfake localization as an important forensic task.

Unlike conventional semantic understanding, deepfake localization requires identifying subtle deviations from normal audio-visual continuity. Although semantic information such as speech content, speaker identity, facial appearance, and scene context provides useful context, accurate localization relies on weak boundary-related signals, including motion discontinuities, acoustic transitions, harmonic variations, and cross-modal synchronization errors. However, these signals are often localized and easily overwhelmed by dominant semantic patterns in multimodal representations, making reliable boundary learning challenging. Recent multimodal large language models (MLLMs) have demonstrated strong multimodal perception and reasoning abilities (Huang et al., 2025; Li et al., 2026). Nevertheless, their pretraining objectives mainly focus on semantic understanding rather than forensic localization. Directly applying MLLMs therefore requires them to implicitly discover weak temporal evidence hidden within high-level semantic representations, which limits precise localization. Figure 1 illustrates this contrast between implicit forensic cue learning and our explicit evidence modeling paradigm.

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.

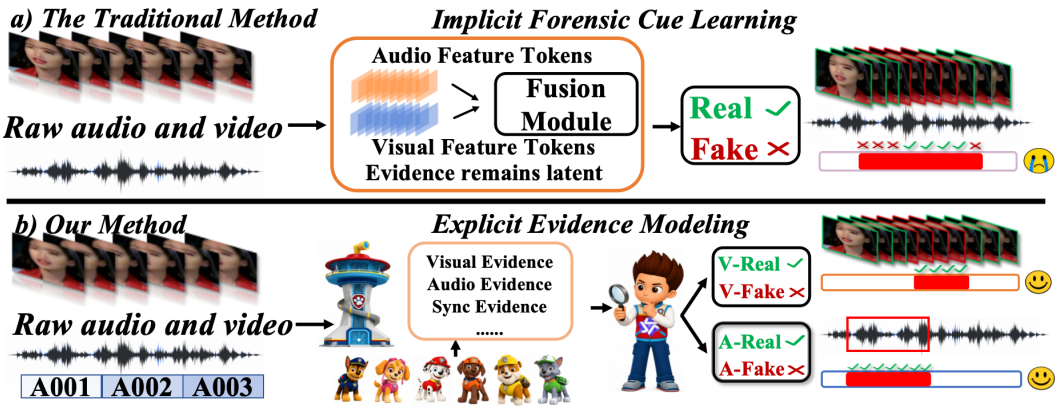

Figure 1: Motivation of EviPatrol. (a) Existing methods implicitly encode forensic evidence cues in latent representations, while (b) Our EviPatrol explicitly models temporally aligned evidence to support accurate localization and evidence-grounded diagnosis.

Existing approaches based on boundary supervision (Heidari et al., 2024), temporal reconstruction (Cai et al., 2023a), and audio-visual discrepancy modeling (Zhang et al., 2023) have achieved considerable progress. However, the localization evidence supporting their predictions is usually embedded in latent features or task-specific scores, lacking an explicit and interpretable representation that can be directly utilized by reasoning modules. This motivates a new paradigm: explicitly exposing boundary-related information before final localization reasoning.

Based on this observation, we propose **EviPatrol**, an explicit boundary modeling framework for fine-grained audio-visual deepfake localization. EviPatrol represents learned localization cues as structured, time-indexed evidence fields, providing explicit temporal references for downstream localization and reasoning. To capture diverse manipulation patterns and mitigate cross-modal decision shortcuts, we combine learnable evidence extraction with modality-aware training.

On a shared audio-visual timeline, a trainable **Boundary-AEP** encoder models abnormality, boundary likelihood, interval support, and cross-modal temporal discrepancies. **BLA-AMR** balances modality-specific evidence learning, while **Safe-AMR** applies training-only consistency regularization to mitigate cross-modal decision shortcuts. Two independently adapted Qwen2.5-Omni (Xu et al., 2025) localizers jointly receive the original audio-visual content and structured evidence to predict audio and visual manipulation intervals, respectively. Finally, a structural verifier checks anchor existence, temporal overlap, modality consistency, and agreement between cited evidence and frozen localization outputs before diagnostic reporting.

Our main contributions are summarized as follows:

- We introduce **explicit boundary modeling** for audio-visual deepfake localization, transforming hidden localization cues into structured, time-indexed evidence representations accessible to downstream reasoning modules.
- We develop a learnable boundary evidence paradigm based on **Boundary-AEP**, together with modality-aware learning strategies including **BLA-AMR** and **Safe-AMR**, to construct adaptive localization evidence and mitigate cross-modal decision shortcuts.
- We propose an evidence-conditioned MLLM localization framework that integrates structured boundary evidence with multimodal reasoning, enabling accurate temporal localization and consistent diagnostic analysis.

2 Related Work

2.1 Audio-Visual Temporal Forgery Localization

Early deepfake detection studies mainly focused on image-level or video-level binary classification, typically producing only global authenticity scores and struggling to localize short manipulated

intervals (Cai et al., 2023a; 2024; Mirsky & Lee, 2021). LAV-DF (Cai et al., 2022) subsequently introduced localized audio, visual, and audio-visual forgery settings. Methods such as BA-TFD (Cai et al., 2022), BA-TFD+ (Cai et al., 2023a), and UMMAFormer (Zhang et al., 2023) improved temporal localization through boundary supervision, temporal anomaly modeling, and cross-modal fusion. However, their forensic cues are mostly embedded in latent features or proposal scores, making them difficult to inspect, verify, and utilize for explanation.

In contrast, EviPatrol organizes boundary cues from audio continuity, visual dynamics, and audio-visual correspondence into explicit temporal evidence, which is further interpreted by the MLLM together with the original audio-visual content for final localization and explanation.

2.2 Modality Bias

Methods such as mDPO (Zhang et al., 2024) alleviate unimodal dominance through modality-conditioned preference optimization or perturbation constraints. However, these methods mainly focus on general response quality and preference alignment, rather than modeling modality-specific learning states and evidence reliability for temporal localization. In audio-visual forgery localization, strong anomalies from one modality may incorrectly influence interval predictions of another modality. EviPatrol therefore separates the evidence, supervision, and prediction responsibilities of audio and visual modalities, and employs BLA-AMR together with training-stage counterfactual consistency constraints to mitigate asymmetric modality learning and cross-modal decision shortcuts.

2.3 Multimodal Large Language Models for Deepfake Analysis

MLLMs provide strong semantic understanding and structured generation capabilities, but accurate temporal localization still requires dedicated temporal modeling (Xie et al., 2026; Li et al., 2026), while forensic MLLMs such as Omni-Fake-R1 (Li et al., 2026) have begun to jointly address deepfake detection, localization, and explanation. However, semantic events usually correspond to clear content or actions, whereas forgery boundaries may only manifest as subtle acoustic changes, discontinuities in visual dynamics, or local audio-visual misalignment. Existing methods also rarely perform separate localization of potentially asynchronous audio and visual forgeries, and lack explicit verification of the temporal and modality consistency of explanation evidence.

EviPatrol provides the MLLM with the original audio, original video, and complete timeline evidence simultaneously. Each predicted interval is associated with corresponding evidence anchors, while conservative boundary fusion and explanation verification ensure consistency among the localized intervals, modality decisions, evidence, and natural-language explanations.

3 Method

3.1 Overview

Given an audio-visual sample $X = (A, V)$, where A and V denote the audio and visual streams, respectively, we study the task of audio-visual deepfake temporal localization and diagnosis. The goal is to identify manipulated modalities and localize the corresponding forged intervals.

For each modality $m \in \{a, v\}$, where a and v denote audio and visual modalities, respectively, the model predicts a set of manipulated segments:

$$\mathcal{S}_m = \{([s_i^m, e_i^m], c_i^m)\}_{i=1}^{N_m}, \tag{1}$$

where N_m denotes the number of predicted intervals, s_i^m and e_i^m represent the start and end timestamps, and c_i^m denotes the confidence score. The audio and visual segments are further combined according to their temporal relationships to obtain the union localization result $\mathcal{S}_u$. Instead of directly predicting video-level manipulation states, EviPatrol derives the final state from modality-specific localization results, ensuring consistency between modality decisions and temporal localization.

Beyond localization, EviPatrol generates an evidence-grounded diagnostic report R, where each prediction is associated with supporting temporal evidence extracted during the localization process. Unlike unconstrained post-hoc explanations, the generated diagnosis is grounded by explicit

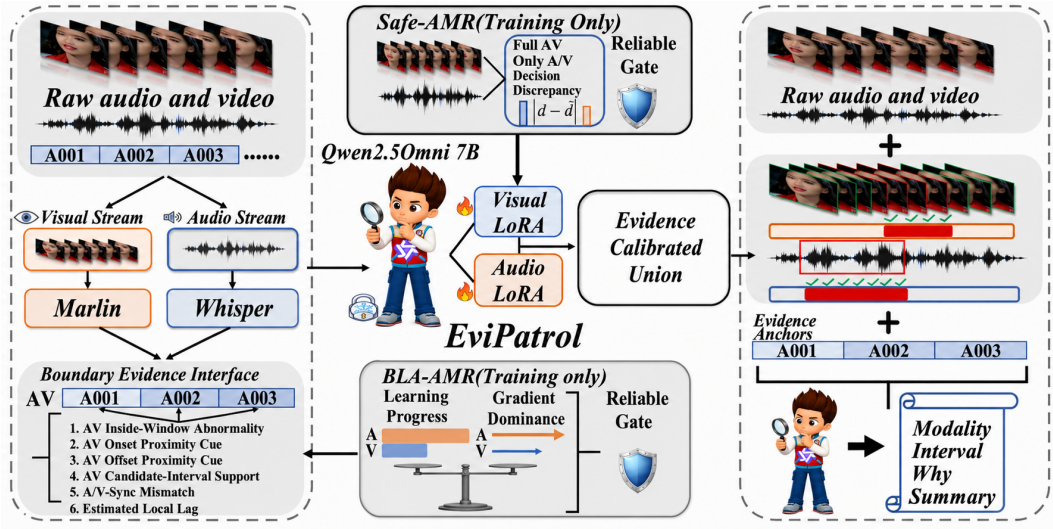

Figure 2: Overview of EviPatrol. Given raw audio-visual inputs, EviPatrol first constructs a boundary evidence interface with structured temporal evidence anchors. Boundary-AEP learns modality-aware boundary evidence, while BLA-AMR and Safe-AMR improve evidence reliability and reduce cross-modal decision shortcuts during training. The generated evidence is integrated with Qwen2.5-Omni based localizers to predict modality-specific forgery intervals and generate evidence-grounded diagnostic summaries.

evidence representations. The key idea of EviPatrol is to introduce a **Boundary Evidence Interface** between audio-visual forensic cue extraction and MLLM reasoning. Instead of directly asking an MLLM to discover subtle manipulation boundaries from raw media, EviPatrol first converts implicit temporal anomalies into structured evidence representations.

As illustrated in Figure 2, EviPatrol uses explicit temporal evidence as an interface between audio-visual forensic cue extraction and MLLM reasoning. The Boundary Evidence Interface transforms fine-grained temporal anomalies into a time-aligned Evidence Pack, while BLA-AMR and Safe-AMR regulate modality learning during training to alleviate asymmetric optimization and cross-modal decision shortcuts. Evidence-Grounded Split-Modal MLLM Localization jointly processes the original audio-visual content and structured evidence to independently localize audio and visual manipulations, followed by evidence-consistent fusion and diagnosis.

3.2 Boundary Evidence Interface

Fine-grained deepfake localization depends on brief and subtle temporal cues that can be obscured by substantially longer authentic content. We therefore introduce a **Boundary Evidence Interface**, implemented by a trainable Boundary-AEP encoder, to transform implicit continuity violations into structured and time-indexed forensic observations. The interface guides the MLLM toward relevant temporal regions but does not directly determine the final forgery prediction.

Given the audio and visual streams A and V, we divide the input into K synchronized temporal anchors:

$$\mathcal{I} = \{I_k = [\tau_k^s, \tau_k^e]\}_{k=1}^{K}, \tag{2}$$

where $\mathcal{I}$ is the complete anchor sequence, I_k denotes the k-th anchor, and τ_k^s and τ_k^e are its start and end timestamps. These anchors provide a shared temporal coordinate system for audio evidence, visual evidence, localization outputs, and diagnostic reports.

We use frozen **MARLIN ViT-B YTF** (Cai et al., 2023b) and **Whisper large-v3** (Radford et al., 2023) encoders to extract dense visual and acoustic features, respectively. Lightweight temporal adapters project both modalities into a shared feature space, and learned local shift alignment models their relative temporal correspondence.

For modality $m \in \{a, v\}$, where a and v denote audio and visual modalities, let $\mathbf{H}_k^m$ be the aligned tokens within anchor I_k. To preserve transitions that may be suppressed by ordinary mean pooling, we construct

$$\mathbf{z}_k^m = \left[\mathrm{Mean}_t(\mathbf{H}_k^m); \mathrm{Mean}_t(\mathbf{H}_{k,R}^m) - \mathrm{Mean}_t(\mathbf{H}_{k,L}^m)\right], \tag{3}$$

where $\mathrm{Mean}_t(\cdot)$ denotes temporal mean pooling, $\mathbf{H}_{k,L}^m$ and $\mathbf{H}_{k,R}^m$ contain the tokens from the first and second halves of the anchor, respectively, and $[\,;\,]$ denotes concatenation. The two terms encode the local state and its intra-anchor transition, respectively.

Boundary-AEP further applies local, multi-scale, and cross-modal temporal modeling to produce modality-specific and relational evidence: For modality $m \in \{a, v\}$, the modality-specific and auxiliary audio-visual evidence are defined as

$$\begin{aligned} \mathbf{e}_k^m &= \left(p_{k,\mathrm{abn}}^m, p_{k,\mathrm{start}}^m, p_{k,\mathrm{end}}^m, p_{k,\mathrm{int}}^m\right), \\ \mathbf{e}_k^{av} &= \left(p_{k,\mathrm{cons}}, \ell_k\right). \end{aligned} \tag{4}$$

Here, $p_{k,\mathrm{abn}}^m$ represents within-anchor continuity abnormality; $p_{k,\mathrm{start}}^m$ and $p_{k,\mathrm{end}}^m$ indicate the proximity of a manipulation onset and offset; and $p_{k,\mathrm{int}}^m$ measures support for the anchor belonging to a coherent candidate interval. The auxiliary variables $p_{k,\mathrm{cons}}$ and ℓ_k represent local audio-visual consistency disagreement and estimated signed temporal lag, respectively. These quantities are intermediate forensic observations rather than final manipulation probabilities.

Finally, a deterministic renderer ρ converts the anchor-level evidence into semantically defined Evidence Cards:

$$c_k = \rho\left(I_k, \mathbf{e}_k^a, \mathbf{e}_k^v, \mathbf{e}_k^{av}\right), \qquad \mathcal{E} = \{c_k\}_{k=1}^K, \tag{5}$$

where c_k is the Evidence Card associated with anchor I_k, and $\mathcal{E}$ is the Evidence Pack covering the complete timeline. Each card preserves the anchor time range and the semantics of its audio, visual, and cross-modal evidence.

Training annotations are used only to supervise Boundary-AEP. At inference time, $\mathcal{E}$ is generated solely from A and V, without access to ground-truth states or manipulation intervals. The downstream MLLM jointly receives $(A, V, \mathcal{E})$, using the evidence to identify regions requiring inspection while retaining the original media as the basis for modality attribution, localization, and diagnosis.

3.3 Evidence-Aware Modality Rebalancing

Audio and visual evidence can exhibit different learning dynamics, causing one modality to dominate joint optimization. We therefore introduce **Boundary-Learnability-Aware Adaptive Modality Rebalancing** (**BLA-AMR**) to regulate modality-specific evidence learning in Boundary-AEP. For modality $m \in \{a, v\}$, where a and v denote audio and visual modalities, respectively, let t denote the training step, $\mathcal{L}_{m,t}^{\mathrm{core}}$ denote the modality-specific core evidence loss at step t, and $\mathcal{L}_{m,0}^{\mathrm{core}}$ denote its warm-up reference. The learning progress is defined as

$$P_{m,t} = \mathrm{clip}_{[0,1]}\left(1 - \mathrm{EMA}\left[\frac{\mathcal{L}_{m,t}^{\mathrm{core}}}{\mathcal{L}_{m,0}^{\mathrm{core}} + \epsilon}\right]\right), \tag{6}$$

where $P_{m,t}$ denotes the learning progress of modality m, $\mathrm{EMA}(\cdot)$ is an exponential moving average, ϵ is a numerical-stability constant, and $\mathrm{clip}_{[0,1]}(\cdot)$ clips its input to $[0, 1]$. BLA-AMR combines the relative progress gap, shared-gradient dominance, and training-time extraction quality to make bounded adjustments to modality-specific evidence losses and sample selection.

We further introduce **Safe-AMR** to reduce cross-modal decision shortcuts during MLLM adaptation. The four states follow the audio–visual order: RR, RF, FR, and FF denote real–real, real–fake, fake–real, and fake–fake, respectively. Therefore, $(\mathrm{RF}, \mathrm{FF})$ and $(\mathrm{RR}, \mathrm{FR})$ share the same visual labels while differing in audio state. Let $\hat{d}_{s,t}$ denote the normalized visual decision loss for state $s \in \{\mathrm{RR}, \mathrm{RF}, \mathrm{FR}, \mathrm{FF}\}$ at training step t. We define the paired-state discrepancies as

$$\begin{aligned} \Delta_{+,t} &= \left[|\hat{d}_{\mathrm{RF},t} - \hat{d}_{\mathrm{FF},t}| - \delta_s\right]_+, \\ \Delta_{-,t} &= \left[|\hat{d}_{\mathrm{RR},t} - \hat{d}_{\mathrm{FR},t}| - \delta_s\right]_+. \end{aligned} \tag{7}$$

where $\Delta_{+,t}$ and $\Delta_{-,t}$ denote the discrepancies for the visual-fake and visual-real pairs, respectively, δ_s is a tolerance margin, and $[x]_+ = \max(x, 0)$. Large discrepancies indicate that visual decisions

vary unnecessarily with the audio condition. Safe-AMR therefore constructs an audio-withheld branch during training, which preserves the visual input, visual evidence, and supervision target while removing raw audio and audio-derived evidence. The normal branch is regularized toward a consistent target-modality decision when the paired branch provides reliable supervision. During inference, the adaptive controllers and auxiliary branch are removed, and the model receives the complete original audio-visual input and Evidence Pack without additional inference cost.

3.4 Evidence-Grounded Split-Modal MLLM Localization

We adopt an MLLM as the final localizer because accurate audio-visual forgery localization requires not only detecting local boundary cues, but also jointly interpreting the original audio, video, and their contextual relationships for modality attribution and continuous interval prediction. Boundary-AEP provides temporally indexed forensic observations, but these anchor-level responses are not final manipulation intervals. We use Qwen2.5-Omni (Xu et al., 2025) to reason over the original media together with the structured evidence.

For sample i, let A_i, V_i, and $\mathcal{E}_i$ denote the original audio, original video, and complete Evidence Pack, respectively. For modality $m \in \{a, v\}$, where a and v denote audio and visual modalities, $\mathcal{E}_i^m$ denotes the modality-focused serialization of $\mathcal{E}_i$. The conditional input and structured output are defined as

$$\begin{aligned} X_i^m &= (\Pi_m, A_i, V_i, \mathcal{E}_i^m), \\ \hat{Y}_i^m &= F_{\Theta,\Delta_m}(X_i^m), \end{aligned} \tag{8}$$

where Π_m is the modality-specific instruction, X_i^m is the complete conditional input, F_{Θ,Δ_m} denotes the adapted Qwen2.5-Omni model, Θ represents the frozen backbone parameters, Δ_m represents the modality-specific LoRA (Hu et al., 2021) parameters, and $\hat{Y}_i^m$ is the generated structured output.

The audio and visual models use the same frozen backbone initialization but independently trained LoRA adapters Δ_a and Δ_v. Thus, *split-modal* refers to the separation of trainable parameters, supervision targets, and prediction responsibilities rather than sensory inputs. Both models retain the complete audio-visual context. The audio model predicts only audio manipulations, whereas the visual model predicts only visual manipulations; evidence from the other modality may provide temporal context but cannot independently justify a manipulation decision for the responsible modality.

For modality m, the generated output is parsed into a set of modality-specific segments:

$$\begin{aligned} \hat{\mathbf{s}}_{ij}^m &= \big(\hat{b}_{ij}^m, \hat{e}_{ij}^m, \hat{c}_{ij}^m, \hat{\mathcal{K}}_{ij}^m\big), \\ \hat{\mathcal{S}}_i^m &= \{\hat{\mathbf{s}}_{ij}^m\}_{j=1}^{N_i^m}, \end{aligned} \tag{9}$$

where j indexes the predicted segments, N_i^m is the number of predicted segments for sample i and modality m, $\hat{b}_{ij}^m$ and $\hat{e}_{ij}^m$ are the continuous start and end timestamps, $\hat{c}_{ij}^m$ is the generated confidence score, and $\hat{\mathcal{K}}_{ij}^m$ is the set of cited Evidence Anchor identifiers. The temporal anchors serve only as inspection coordinates; the predicted boundaries are not restricted to anchor endpoints.

To maintain consistency between modality decisions and localization, we adopt a *segments-only* protocol. The state of modality m is derived as

$$\hat{z}_i^m = \mathbf{1}\Big[\Big|\mathrm{Valid}\big(\hat{\mathcal{S}}_i^m\big)\Big| > 0\Big], \tag{10}$$

where $\mathrm{Valid}(\cdot)$ filters segments that violate the output schema or temporal constraints, $|\cdot|$ denotes set cardinality, and $\mathbf{1}[\cdot]$ is the indicator function. The value $\hat{z}_i^m = 1$ indicates that modality m is predicted as manipulated. The global state is deterministically reconstructed from $(\hat{z}_i^a, \hat{z}_i^v)$ in audio–visual order.

After obtaining the modality-specific audio and visual intervals, we construct their temporal union based on overlap and calibrate each union interval's confidence using Boundary Evidence. This step only re-scores existing intervals; it does not modify their boundaries, modality states, or number.

3.5 Evidence-Grounded Diagnostic Reporting

After localization and confidence calibration, EviPatrol freezes all modality decisions, temporal intervals, confidence scores, and evidence references. An explanation adapter then conditions on the original audio-visual input, the complete Evidence Pack, and the frozen localization result to generate concise diagnostic text. During training, the localization parameters and outputs remain fixed, and only the explanation continuation is optimized. The explanation stage therefore describes an existing prediction without modifying its modality state, temporal boundaries, or confidence.

For sample i, modality $m \in \{a, v\}$, and its j-th predicted interval, let $\mathbf{r}_{ij}^m$ denote the corresponding interval-level diagnostic rationale. We construct it as

$$\mathbf{r}_{ij}^m = \boldsymbol{\pi}_{ij}^m \oplus \mathbf{p}_{ij}^m, \tag{11}$$

where $\boldsymbol{\pi}_{ij}^m$ denotes deterministic provenance derived from the frozen localization result and Evidence Pack, including the responsible modality, temporal interval, overlapping Evidence Anchors, and available evidence types. $\mathbf{p}_{ij}^m$ denotes the model-generated observation and diagnostic conclusion, and $\oplus$ denotes deterministic assembly. The model is not allowed to generate interval endpoints, anchor identifiers, or evidence-field names. Before presentation, a verifier checks anchor existence, temporal overlap, modality consistency, and agreement with the frozen prediction. Verified interval-level rationales are aggregated into a video-level diagnostic report R_i. If verification fails, the detailed rationale is withheld while the complete localization result is retained.

3.6 Training Objectives and Inference Protocol

EviPatrol is optimized stage by stage rather than through a single end-to-end objective. Let Θ_{shift} and Θ_{AEP} denote the parameters of the shift-alignment module and Boundary-AEP, respectively; let Δ_a and Δ_v denote the audio and visual LoRA parameters; ϕ_{prop} denote the proposal-quality scorer; and Δ_{exp} denote the explanation-adapter parameters. The training order is

$$\Theta_{\text{shift}} \rightarrow \Theta_{\text{AEP}} \rightarrow \{\Delta_a, \Delta_v\} \rightarrow \phi_{\text{prop}} \rightarrow \Delta_{\text{exp}}, \tag{12}$$

where each upstream component is frozen before training the subsequent stage. The audio and visual adapters are independently initialized and optimized from the same frozen Qwen2.5-Omni backbone.

Boundary-AEP is trained with a multi-task evidence objective:

$$\mathcal{L}_{\text{AEP}} = \mathcal{L}_{\text{abn}} + \mathcal{L}_{\text{bnd}} + \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{recon}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{lag}}\mathcal{L}_{\text{lag}}, \tag{13}$$

where the first three terms supervise abnormality, onset/offset boundaries, and candidate-interval support; the remaining terms model temporal reconstruction, audio-visual consistency disagreement, and local lag. BLA-AMR and Safe-AMR are enabled only during training. The coefficients λ_{cons} and λ_{lag} balance the corresponding auxiliary losses.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate EviPatrol on LAV-DF and AV-Deepfake1M. LAV-DF serves as a standard benchmark for comparison with existing temporal forgery localization methods, whereas AV-Deepfake1M evaluates the framework in a substantially larger and more diverse setting. LAV-DF contains 136,304 clips with localized audio, visual, and joint audio-visual manipulations. We follow the official split of 78,703 training, 31,501 development, and 26,100 test clips. AV-Deepfake1M contains 1,146,760 videos from more than 2,000 identities, including 286,721 real and 860,039 manipulated samples. We use its official split of 746,180 training, 57,340 development, and 343,240 test videos. For both datasets, we report AP for audio, visual, and union localization results under standard temporal-IoU thresholds.

Implementation Details. We use frozen MARLIN ViT-B YTF (Cai et al., 2023b) and Whisper large-v3 (Radford et al., 2023) encoders to extract dense visual and acoustic representations, respectively. Videos are processed at 25 fps, and audio is resampled to 16 kHz. The shared evidence timeline uses a nominal anchor resolution of 0.5 seconds, with at most 32 anchors per clip. Boundary-AEP employs a 512-dimensional hidden space, a four-layer temporal encoder, and a three-level

Table 1: Comparison with existing methods on LAV-DF. All results are reported as AP (%). Average denotes the mean AP over AP@0.50, AP@0.75, and AP@0.95. Evidence-Only uses only extracted evidence without original audio-visual inputs, while Visual Head and Audio Head denote modality-specific localization branches with complete audio-visual context. Full denotes the complete evidence-grounded framework. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	Modality	AP@ 0.50	AP@ 0.75	AP@ 0.95	Average
BMN (Lin et al., 2019)	CVPR'19	V	24.01	07.61	00.07	10.56
MDS (Chugh et al., 2020)	ACM MM'20	AV	12.8	01.60	0.00	04.80
ActionFormer (Zhang et al., 2022)	ECCV'22	V	77.23	51.11	06.96	45.10
TriDet (Shi et al., 2023)	CVPR'23	V	87.43	73.15	07.71	56.10
BA-TFD (Cai et al., 2022)	DICTA'22	AV	78.93	40.01	00.19	39.71
BA-TFD+ (Cai et al., 2023a)	CVIU'23	AV	96.30	84.96	04.44	61.90
TE-TAD (Kim et al., 2024)	CVPR'24	V	83.4	63.5	08.14	51.68
TadTR (Liu et al., 2022)	CVPR'22	AV	80.2	61.1	05.20	48.83
Intra-/Cross-modal Sync (Anshul et al., 2025)	ICCV'25	AV	87.40	66.80	05.72	53.30
DeformTrace (Zhu et al., 2026)	AAAI'26	AV	96.16	89.29	34.23	73.22
EviPatrol (Prompt-Only)	–	AV	06.07	02.11	00.00	02.72
EviPatrol (SFT-Only)	–	AV	42.17	32.68	22.09	32.31
EviPatrol (Evidence-Only)	–	AV	48.42	12.91	00.03	20.45
EviPatrol (Visual-Head)	–	AV	87.45	77.86	71.27	78.86
EviPatrol (Audio-Head)	–	AV	89.43	86.90	84.99	87.11
EviPatrol (Full)	–	AV	**97.98**	**95.59**	**86.17**	**93.25**

Table 2: Comparison with existing methods on AV-Deepfake1M. All results are reported as AP (%). Average denotes the mean AP over AP@0.50, AP@0.75, AP@0.90, and AP@0.95. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Venue	Modality	AP@ 0.50	AP@ 0.75	AP@ 0.90	AP@ 0.95	Average
ActionFormer (Zhang et al., 2022)	ECCV'22	V	21.25	07.77	01.68	00.06	07.69
TriDet (Shi et al., 2023)	CVPR'23	V	27.51	07.64	00.78	00.08	09.00
BA-TFD (Cai et al., 2022)	DICTA'22	AV	37.37	06.34	00.19	00.02	10.98
BA-TFD+ (Cai et al., 2023a)	CVIU'23	AV	44.42	13.64	00.46	00.03	14.64
UMMAFormer (Zhang et al., 2023)	ACM MM'23	AV	51.64	28.07	04.65	01.58	21.49
Intra-/Cross-modal Sync (Anshul et al., 2025)	ICCV'25	AV	29.17	10.76	01.24	00.16	10.33
DeformTrace (Zhu et al., 2026)	AAAI'26	AV	**88.16**	74.29	24.23	08.48	48.79
EviPatrol (Visual-Head)	–	AV	79.74	76.87	57.89	43.29	64.45
EviPatrol (Audio-Head)	–	AV	86.45	82.98	61.44	46.87	69.44
EviPatrol (Full)	–	AV	87.45	**84.34**	**66.98**	**49.87**	**72.16**

temporal pyramid. The shift-alignment module is trained only on real–real training clips. The modality-specific localizers are instantiated from identical frozen Qwen2.5-Omni-7B (Xu et al., 2025) backbone weights and equipped with independently trained audio and visual LoRA adapters. Both adapters use rank 8, scaling factor 16, and dropout 0.05. All trainable parameters are optimized using AdamW with bfloat16 mixed precision. Safe-AMR is instantiated through an audio-withheld branch during visual adaptation, while the audio adapter uses fixed state weighting. Unless otherwise specified, full-scale experiments are conducted on eight NVIDIA A800 80GB GPUs. Additional optimization schedules, checkpoint-selection rules, and calibration settings are provided in the appendix.

Table 3: Component ablation of EviPatrol on LAV-DF. B-AEP denotes Boundary-AEP, B-AMR denotes BLA-AMR, and S-AMR denotes Safe-AMR. A-mAP, V-mAP, and U-mAP denote audio, visual, and union localization mAP, respectively, averaged over IoU thresholds $\{0.50, 0.75, 0.95\}$.

#	B-AEP	B-AMR	S-AMR	MLLM	A-mAP	V-mAP	U-mAP
1	√	×	×	×	–	–	21.74
2	×	×	×	√	44.76	18.56	32.31
3	√	×	×	√	72.92	51.54	63.25
4	√	√	×	√	<u>80.23</u>	68.72	74.28
5	√	×	√	√	78.97	<u>70.42</u>	<u>83.78</u>
6	√	√	√	√	**87.11**	**78.86**	**93.25**

4.2 Comparison with Existing Methods

We compare EviPatrol with representative temporal localization and audio-visual deepfake localization methods on LAV-DF and AV-Deepfake1M, as shown in Tables 1 and 2.

Generic temporal localization methods, such as ActionFormer and TE-TAD, perform reasonably under loose IoU thresholds but degrade substantially under stricter boundary evaluation, indicating difficulty in capturing subtle forgery transitions. Audio-visual deepfake localization methods, including BA-TFD+ and Intra-/Cross-modal Synchronization, improve localization through multimodal representations, boundary modeling, or cross-modal consistency, but still show limited precision at high IoU thresholds.

EviPatrol achieves consistent improvements on both benchmarks. On LAV-DF, it obtains 97.98% AP@0.50, 95.59% AP@0.75, and 86.17% AP@0.95, with particularly large gains under strict localization criteria. On AV-Deepfake1M, EviPatrol reaches 49.87% AP@0.95, demonstrating strong performance in a larger and more diverse setting. The consistent gains at high IoU thresholds support the effectiveness of explicit temporal evidence modeling for precise manipulation boundary localization. EviPatrol also demonstrates stronger cross-dataset generalization than the compared methods, with additional generalization results and analyses provided in the appendix.

4.3 Ablation Studies

Table 1 compares different EviPatrol variants on LAV-DF. Prompt-Only and SFT-Only perform substantially worse without explicit evidence modeling. Evidence-Only, which provides the MLLM with structured evidence but removes the original audio-visual input, also falls behind Full, highlighting the complementary roles of explicit evidence and raw media. Full further improves over the individual Audio-Head and Visual-Head variants.

Table 3 further examines the contribution of each component. Removing Boundary-AEP consistently degrades audio, visual, and union localization. BLA-AMR improves both modalities, with a larger gain on visual localization, while Safe-AMR brings additional improvements by mitigating cross-modal decision shortcuts. Together, these results demonstrate the complementary effects of explicit evidence modeling and modality-aware training. Additional ablations and analyses are provided in the appendix.

5 Conclusion

In this paper, we propose EviPatrol, an evidence-grounded MLLM framework for fine-grained audio-visual deepfake localization and diagnosis. EviPatrol addresses the limitation of implicit forensic cue learning by introducing explicit boundary evidence modeling, which provides temporally aligned evidence for MLLM reasoning. We further mitigate modality imbalance and cross-modal shortcuts through evidence-aware modality optimization, enabling reliable modality-specific localization. Beyond localization, EviPatrol supports evidence-grounded diagnosis by linking each prediction to its corresponding temporal observations. Extensive experiments on LAV-DF and AV-Deepfake1M demonstrate the effectiveness of EviPatrol under strict localization criteria.

AI USE STATEMENT

Generative AI tools were used only to assist with language polishing and improving the clarity of the manuscript. All technical content, experimental design, analysis, results, and conclusions were developed and verified by the authors. The authors reviewed and approved all AI-assisted revisions.

ETHICS STATEMENT

EviPatrol is intended to assist audio-visual media verification by localizing suspected manipulations and providing evidence-grounded diagnoses. Its predictions and explanations may be imperfect and should therefore support, rather than replace, human judgment. Real-world use may also involve risks related to false attribution, privacy, fairness, and dual use, especially when forensic evidence is applied beyond the evaluated benchmarks. Appropriate human oversight and safeguards are therefore important in deployment.

The human evaluation involved 30 participants, as described in Appendix E.3. Participation was voluntary, and participants provided informed consent. Participant information was handled confidentially, and results are reported only in aggregate.

REPRODUCIBILITY STATEMENT

The main paper describes the overall architecture, pretrained backbones, and principal adaptation settings. Appendix A specifies the task definitions, evaluation protocols, and output processing. Appendix B details temporal alignment, Boundary-AEP, and Evidence Pack construction. Appendix C provides localization prompts, input/output examples, parsing rules, deterministic union construction, and the diagnosis procedure. Appendix D describes the BLA-AMR and Safe-AMR objectives, while Appendix F records implementation and serialization conventions. The anonymous supplementary code provides evidence schemas, evidence-rendering utilities, illustrative prompts, output parsing, and deterministic interval union, accompanied by synthetic examples and interface tests.

REFERENCES

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, October 2025.

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 *2022 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, 2023a. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2023.103818. URL https://www.sciencedirect.com/science/article/pii/S1077314223001984.

Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning, 2023b. URL https://arxiv.org/abs/2211.06627.

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*, MM '24, pp. 7414–7423, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680795. URL https://doi.org/10.1145/3664647.3680795.

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, MM '20, pp. 439–447, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413700. URL https://doi.org/10.1145/3394171.3413700.

Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery*, 14(2):e1520, 2024. doi: https://doi.org/10.1002/widm.1520. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1520.

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, 2021. URL https://arxiv.org/abs/2106.09685.

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, June 2025.

Ho-Joong Kim, Jung-Ho Hong, Heejo Kong, and Seong-Whan Lee. Te-tad: Towards full end-to-end temporal action detection via time-aligned coordinate expression, 2024. URL https://arxiv.org/abs/2404.02405.

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

Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation, 2019. URL https://arxiv.org/abs/1907.09702.

Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441, 2022. ISSN 1941-0042. doi: 10.1109/tip.2022.3195321. URL http://dx.doi.org/10.1109/TIP.2022.3195321.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), January 2021. ISSN 0360-0300. doi: 10.1145/3425780. URL https://doi.org/10.1145/3425780.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling, 2023. URL https://arxiv.org/abs/2303.07347.

Shichuang Xie, Tong Qiao, Sheng Li, Xinpeng Zhang, Jiantao Zhou, and Guorui Feng. Deepfake detection in the aigc era: A survey, benchmarks, and future perspectives. *Information Fusion*, 127:103740, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.103740. URL https://www.sciencedirect.com/science/article/pii/S1566253525008024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers, 2022. URL https://arxiv.org/abs/2202.07925.

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*, MM '23, pp. 8749–8759, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3613767. URL https://doi.org/10.1145/3581783.3613767.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward, 2024. URL https://arxiv.org/abs/2404.01258.

Xiaodong Zhu, Suting Wang, Yuanming Zheng, Junqi Yang, Yangxu Liao, Yuhong Yang, Weiping Tu, and Zhongyuan Wang. Deformtrace: A deformable state space model with relay tokens for temporal forgery localization, 2026. URL https://arxiv.org/abs/2603.04882.

A TASK AND EVALUATION PROTOCOL

This section defines the task, dataset usage, evaluation metrics, and output-audit protocol used throughout the paper. Dataset statistics and the split evaluated in each experiment are reported separately. Unless otherwise stated, development-set results are computed on the complete official development split. Reporting the size of an official test set does not imply that the test set was accessed.

A.1 TASK DEFINITION AND OUTPUT PROTOCOL

Let $x_i = (A_i, V_i)$ denote the i-th audio–visual clip with duration T_i, where A_i and V_i are its audio and visual streams, respectively. For modality $m \in \{a, v\}$, where a and v denote audio and visual modalities, the ground-truth manipulated intervals are

$$\mathcal{G}_i^m = \{[b_{ir}^m, e_{ir}^m]\}_{r=1}^{M_i^m}, \qquad 0 \le b_{ir}^m < e_{ir}^m \le T_i, \tag{14}$$

where r indexes the ground-truth intervals, M_i^m is the number of ground-truth intervals for modality m, and b_{ir}^m and e_{ir}^m are the corresponding start and end timestamps. An empty set indicates that modality m is authentic.

The audio and visual localization branches return raw segment sets

$$\widehat{\mathcal{S}}_i^m = \left\{ \left(\hat{b}_{ij}^m, \hat{e}_{ij}^m, \hat{c}_{ij}^m, \widehat{\mathcal{K}}_{ij}^m \right) \right\}_{j=1}^{N_i^m}, \tag{15}$$

where j indexes the raw predictions, N_i^m is the number of raw predicted intervals, $\hat{b}_{ij}^m$ and $\hat{e}_{ij}^m$ are their start and end timestamps in seconds, $\hat{c}_{ij}^m$ is the raw confidence value, and $\widehat{\mathcal{K}}_{ij}^m$ is an optional set of cited Evidence Card anchor IDs. Anchor citations support auditing and diagnosis but do not affect temporal-overlap computation.

After deterministic parsing, all segment records pass through the same validation and sanitization procedure. The resulting valid segment set is

$$\overline{\mathcal{S}}_i^m = \mathrm{Valid}\left(\widehat{\mathcal{S}}_i^m; T_i, \theta_m \right), \tag{16}$$

where $\mathrm{Valid}(\cdot)$ denotes the deterministic validation and sanitization operator, and θ_m is the confidence threshold for modality m, if confidence filtering is used. All thresholds are selected exclusively on the permitted source development split and are frozen before evaluation. When no confidence filtering is applied, $\theta_m = 0$.

The predicted state of each modality is derived only from whether its valid interval set is empty:

$$\hat{z}_i^m = \mathbf{1}\left[\left| \overline{\mathcal{S}}_i^m \right| > 0 \right], \qquad m \in \{a, v\}, \tag{17}$$

where $\mathbf{1}[\cdot]$ is the indicator function, $|\cdot|$ denotes set cardinality, and $\hat{z}_i^m = 0$ and $\hat{z}_i^m = 1$ indicate real and fake, respectively.

The four global states follow a fixed audio–visual order:

$(\hat{z}_i^a, \hat{z}_i^v)$	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
State	RR	RF	FR	FF.

(18)

Here, R and F denote real and fake, respectively, while the first and second letters refer to the audio and visual modalities. A generated real/fake token is ignored during evaluation and cannot override the segments-only decision in either direction. A modality is evaluated as fake if and only if its responsible branch produces at least one valid interval.

Let $\mathcal{G}_i^u$ denote the official union ground-truth annotation for clip i, where u denotes union localization. The predicted union is constructed deterministically from the sanitized audio and visual outputs:

$$\widehat{\mathcal{U}}_i = \mathrm{Merge}\left(\overline{\mathcal{S}}_i^a \cup \overline{\mathcal{S}}_i^v \right), \tag{19}$$

where $\mathrm{Merge}(\cdot)$ returns connected components of overlapping or exactly touching predicted temporal support. Each merged component uses the temporal envelope of its contributing intervals. Its

direct-union confidence is initialized as the maximum confidence among its contributing intervals, and its evidence provenance is the union of their anchor-ID sets. Union localization is therefore a deterministic composition of the audio and visual predictions rather than a separately learned third localizer. The Merge operation is not applied to official union ground-truth annotations.

For each dataset, the use of a learned confidence-only scorer is determined exclusively on the permitted development split and is frozen before final evaluation. If selected, the scorer may update only the confidence of an existing union interval. It cannot add or remove intervals, move their endpoints, change modality states, alter interval count, or modify evidence provenance. Otherwise, the direct-union confidence is retained without learned calibration.

A.2 Datasets and Evaluation Settings

LAV-DF contains 136,304 clips with localized audio, visual, and joint audio–visual manipulations. AV-Deepfake1M contains 1,146,760 videos from more than 2,000 identities, including 286,721 real and 860,039 manipulated samples. AV-Deepfake1M officially refers to its development split as the validation split; we use "development" as the common term below.

For LAV-DF, the official `fake_periods` are used directly as the union ground truth $\mathcal{G}_i^u$, without additional temporal merging. LAV-DF provides modality-level audio and visual manipulation labels but does not provide separate modality-specific temporal spans. For modality-specific analysis, the official `fake_periods` are assigned to each modality marked as manipulated in the metadata, while an unmanipulated modality receives an empty interval set. When both modalities are marked as manipulated, the same official periods are assigned to both modality-specific sets. These derived audio and visual annotations are used only for modality-specific diagnostic evaluation. Comparisons under the official LAV-DF union protocol always use the original `fake_periods`.

For AV-Deepfake1M, the official audio, visual, and union temporal annotations are used directly for $\mathcal{G}_i^a$, $\mathcal{G}_i^v$, and $\mathcal{G}_i^u$, respectively.

For in-domain experiments, we follow the official dataset partitions. Model optimization uses the training split, while checkpoint selection, hyperparameter tuning, prompt configuration, and confidence calibration, when applicable, are performed on the development split. All model and evaluation settings are fixed before test evaluation.

Our LAV-DF→AV-Deepfake1M cross-dataset experiment follows a strict zero-shot protocol. The model is trained on the complete official LAV-DF training split. All checkpoint, prompt, threshold, calibration, and post-processing decisions are restricted to the official LAV-DF development split. The resulting source-selected model is frozen before target-domain inference.

The frozen model is evaluated on the official AV-Deepfake1M test split. The model is directly transferred from LAV-DF to AV-Deepfake1M without target-domain adaptation.

Target-domain labels are used only by the evaluator after all predictions have been generated and frozen. They are never exposed to the model and are not used for adaptation, checkpoint selection, prompt design, threshold tuning, confidence calibration, or post-processing. All methods reported in the cross-dataset comparison are evaluated on the same official AV-Deepfake1M test split using the same canonical interval schema, output audit, and evaluator.

A.3 Metrics and Output Audit

For a predicted interval $p = [b_p, e_p]$ and a ground-truth interval $g = [b_g, e_g]$, where b_p and b_g denote their start timestamps and e_p and e_g denote their end timestamps, temporal intersection over union is

$$\mathrm{tIoU}(p, g) = \frac{|p \cap g|}{|p| + |g| - |p \cap g|}, \tag{20}$$

where $|p|$, $|g|$, and $|p \cap g|$ denote the temporal durations of the predicted interval, ground-truth interval, and their intersection, respectively.

We follow the benchmark one-dimensional temporal detection protocol. At temporal-IoU threshold τ, predictions from the requested output stream are sorted globally by confidence. A prediction can match only an unmatched ground-truth interval from the same record and the same output stream. It

is counted as a true positive if the highest eligible tIoU is at least τ; otherwise, it is counted as a false positive. Matching is greedy and one-to-one. AP@τ is computed from the resulting precision–recall curve using the evaluator fixed for the corresponding experiment.

We use the protocol-specific threshold sets

$$\begin{aligned} \mathcal{T}_{\mathrm{LAV}} &= \{0.50, 0.75, 0.95\}, \\ \mathcal{T}_{\mathrm{AV1M}} &= \{0.50, 0.75, 0.90, 0.95\}, \\ \mathcal{T}_{\mathrm{cross}} &= \{0.50, 0.75, 0.95\}. \end{aligned} \tag{21}$$

For evaluation protocol $\pi \in \{\mathrm{LAV}, \mathrm{AV1M}, \mathrm{cross}\}$, $\mathcal{T}_\pi$ denotes its corresponding set of temporal-IoU thresholds.

For evaluation protocol π and output stream $s \in \{a, v, u\}$, the reported summary metric is

$$\mathrm{mAP}^s_\pi = \frac{1}{|\mathcal{T}_\pi|} \sum_{\tau \in \mathcal{T}_\pi} \mathrm{AP}^s \,@\tau, \tag{22}$$

where a, v, and u denote audio, visual, and union localization, respectively, and $\mathrm{AP}^s \,@\tau$ is the average precision of output stream s at temporal-IoU threshold τ. The corresponding summary quantities are denoted A-mAP, V-mAP, and U-mAP. The "Average" column in each result table is the same arithmetic mean over the AP thresholds displayed for that evaluation protocol and output stream. No unreported threshold is included.

Evaluation is preceded by a deterministic, fail-closed output audit. Every prediction is first bound to its input record ID, and the evaluator requires exactly one prediction object for every record in the complete evaluation split. Duplicate IDs, missing IDs, extra IDs, or any mismatch between the prediction-ID set and the official evaluation-ID set cause the run to fail the audit; the evaluation denominator is never reduced.

For EviPatrol's generative outputs, the parser first attempts to extract the first answer JSON object. A fixed compact-prefix recovery rule may recover localization fields from an otherwise truncated continuation. If neither parsing route succeeds, the corresponding record remains in the evaluation with empty audio and visual predictions, and the parser-failure counter is incremented. Predictions from non-generative baselines are converted from their native interval format into the same canonical segment schema before the common validation and evaluation steps.

Within a parsed EviPatrol output, malformed segment entries are handled only by the predeclared sanitizer. A reversed pair of finite endpoints is canonicalized once, endpoints are clipped to $[0, T_i]$, and zero- or negative-duration intervals are removed. Entries with non-finite endpoints or non-finite confidence values are rejected and counted. Finite confidence values are clipped to $[0, 1]$. Any confidence threshold is frozen before evaluation and may be selected only on the permitted source development split.

After sanitization, Eq. (17) reconstructs the modality and global states, while Eq. (19) constructs the deterministic union. Malformed outputs therefore cannot disappear from metric accounting: a total parse failure contributes an empty prediction for that record, whereas a valid but poorly localized interval contributes according to the standard false-positive and false-negative rules.

B Boundary Evidence Interface

This section describes how EviPatrol converts raw audio and video into the structured boundary evidence consumed by the downstream MLLM. Boundary-AEP produces temporally indexed forensic observations rather than final manipulation intervals or clip-level decisions.

B.1 Temporal Anchors and Dense Feature Extraction

Let $x = (A, V)$ denote an audio–visual clip of duration $T > 0$, where A and V are its audio and visual streams, respectively. We use a nominal anchor width of $\Delta_0 = 0.5$ seconds and a maximum anchor count of $K_{\max} = 32$. Define

$$K_0 = \left\lceil \frac{T}{\Delta_0} \right\rceil, \qquad K = \min(K_0, K_{\max}), \tag{23}$$

where K_0 is the number of anchors required at the nominal resolution and K is the number of valid anchors retained for the clip.

The ordered anchor sequence is

$$\mathcal{I} = \langle I_1, I_2, \ldots, I_K \rangle, \qquad I_k = [\tau_k^s, \tau_k^e], \tag{24}$$

where $k \in \{1, \ldots, K\}$ indexes the anchors, and τ_k^s and τ_k^e are the start and end timestamps of anchor I_k.

When $K_0 \leq K_{\max}$, the anchor boundaries are

$$\tau_k^s = (k-1)\Delta_0, \qquad \tau_k^e = \min(k\Delta_0, T). \tag{25}$$

Thus, all anchors use the nominal width except that the final anchor may be clipped to the clip duration.

When $K_0 > K_{\max}$, the complete interval $[0, T]$ is divided into $K_{\max}$ duration-adaptive windows:

$$\tau_k^s = \frac{(k-1)T}{K_{\max}}, \qquad \tau_k^e = \frac{kT}{K_{\max}}, \qquad K = K_{\max}. \tag{26}$$

This construction preserves coverage of the complete timeline rather than truncating content beyond a fixed duration. Each anchor receives a persistent identifier id_k, which is shared by feature extraction, Evidence Cards, localization outputs, and diagnostic reports.

Frozen MARLIN ViT-B YTF and Whisper large-v3 encoders extract dense visual and acoustic representations, respectively. Only Whisper encoder states are used; no ASR transcript is provided to Boundary-AEP. Lightweight temporal adapters project both modalities into a shared d_h-dimensional feature space, where $d_h = 512$ in all experiments, while preserving their temporal coordinates.

A separately trained shift-alignment adapter models local audio–visual temporal correspondence. It is trained on RR clips from the corresponding training split using controlled synthetic audio shifts and their known shift targets. No forgery-interval annotation is used during this stage. The shift-alignment adapter is frozen before Boundary-AEP training and remains frozen during inference. For a segment length τ on the dense feature grid, the synthetic shift classes are $d \in \{-(\tau-1), \ldots, 0, \ldots, \tau-1\}$. The index convention is $\widetilde{\mathbf{a}}_t = \mathbf{a}_{t+d}$: a positive class selects a later source-audio feature for the current visual position. The shift is converted to seconds using the dense feature-grid spacing, not the possibly adaptive anchor width. Boundary segments use only shifts with available context; padded positions are not treated as observed audio.

For modality $m \in \{a, v\}$, where a and v denote audio and visual modalities, let

$$\mathbf{H}_k^m \in \mathbb{R}^{L_k^m \times d_h} \tag{27}$$

denote the aligned dense tokens inside anchor I_k, where L_k^m is the number of valid tokens for modality m in that anchor. The token sequence is deterministically divided at the temporal midpoint of the anchor into its first and second halves, $\mathbf{H}_{k,L}^m$ and $\mathbf{H}_{k,R}^m$, respectively.

Boundary-preserving pooling retains both the average anchor content and its internal temporal transition:

$$\mathbf{z}_k^m = \left[\mathrm{Mean}_t\left(\mathbf{H}_k^m\right); \mathrm{Mean}_t\left(\mathbf{H}_{k,R}^m\right) - \mathrm{Mean}_t\left(\mathbf{H}_{k,L}^m\right)\right], \tag{28}$$

where $\mathrm{Mean}_t(\cdot)$ denotes mean pooling over the temporal-token dimension and $[\,;\,]$ denotes feature concatenation. The first term describes the local anchor state, while the second term describes its intra-anchor transition. This representation preserves short temporal changes that ordinary mean pooling may suppress.

B.2 Boundary-AEP Architecture and Evidence Contract

To include inter-anchor variation, we first define

$$\Delta\mathbf{z}_k^m = \begin{cases} \mathbf{0}, & k = 1, \\ \mathbf{z}_k^m - \mathbf{z}_{k-1}^m, & k > 1, \end{cases} \tag{29}$$

where $\mathbf{0}$ is a zero vector with the same dimension as $\mathbf{z}_k^m$. The modality-specific anchor representation is then

$$\mathbf{u}_k^m = P_m\left([\mathbf{z}_k^m; \Delta\mathbf{z}_k^m]\right), \tag{30}$$

where $P_m : \mathbb{R}^{4d_h} \to \mathbb{R}^{d_h}$ is a trainable modality-specific projection. The input dimension is $4d_h$ because both $\mathbf{z}_k^m$ and $\Delta\mathbf{z}_k^m$ contain two d_h-dimensional terms.

Boundary-AEP also models whether the current anchor can be reconstructed from its surrounding temporal context. Let h denote the fixed center-excluded neighborhood radius and define

$$\mathcal{N}_k^m = \left\langle \mathbf{u}_j^m \,\middle|\, 1 \le j \le K,\ 0 < |j-k| \le h \right\rangle, \tag{31}$$

where $\mathcal{N}_k^m$ contains only valid neighboring anchors and excludes the current anchor k. A modality-specific reconstruction module

$$F_{\mathrm{rec}}^m : \mathcal{N}_k^m \to \mathbb{R}^{d_h}$$

predicts the current anchor representation from this neighborhood. The reference reconstruction block first takes the mask-normalized mean of the valid, center-excluded neighbors and then applies a two-layer MLP with a GELU activation. The center feature is not an input to that MLP. If no neighboring observation is available, the pooled context is zero; invalid neighbors never contribute to the normalization. The resulting reconstruction residual is

$$\varepsilon_{k,\mathrm{rec}}^m = \frac{1}{d_h} \left\| \mathbf{u}_k^m - F_{\mathrm{rec}}^m\left(\mathcal{N}_k^m\right) \right\|_1, \tag{32}$$

where $\|\cdot\|_1$ is the L_1 norm. A large $\varepsilon_{k,\mathrm{rec}}^m$ indicates that the current anchor is difficult to explain from its surrounding temporal context. This residual is an internal continuity observation and must not be interpreted as a learned reliability score.

The anchor representations, reconstruction information, and cross-modal relations are processed by local temporal operators, a four-layer temporal encoder, cross-modal interaction, and a three-level temporal pyramid. The local operators are residual, depthwise temporal convolutions followed by pointwise channel mixing. The pyramid uses strided temporal convolutions, projects each level, resizes the levels back to the anchor grid, and concatenates them before the output projection. Thus, all evidence heads remain aligned with the original anchor identifiers rather than producing a second timeline. For each valid anchor, Boundary-AEP produces the modality-specific evidence

$$\mathbf{e}_k^m = \left(p_{k,\mathrm{abn}}^m, p_{k,\mathrm{start}}^m, p_{k,\mathrm{end}}^m, p_{k,\mathrm{int}}^m\right), \qquad m \in \{a, v\}, \tag{33}$$

and the auxiliary audio–visual evidence

$$\mathbf{e}_k^{av} = \left(p_{k,\mathrm{cons}}, \ell_k\right). \tag{34}$$

The probability-like scores

$$p_{k,\mathrm{abn}}^m, \quad p_{k,\mathrm{start}}^m, \quad p_{k,\mathrm{end}}^m, \quad p_{k,\mathrm{int}}^m, \quad p_{k,\mathrm{cons}}$$

lie in $[0, 1]$. Specifically, $p_{k,\mathrm{abn}}^m$ measures local continuity abnormality; $p_{k,\mathrm{start}}^m$ and $p_{k,\mathrm{end}}^m$ indicate proximity to a manipulation onset and offset; and $p_{k,\mathrm{int}}^m$ measures support for anchor I_k belonging to a coherent candidate interval.

The cross-modal score $p_{k,\mathrm{cons}}$ follows the notation of the main paper and measures a local audio–visual consistency violation: larger values indicate stronger disagreement rather than stronger agreement. The variable $\ell_k \in \mathbb{R}$ is the estimated signed audio–visual temporal lag in seconds under the fixed convention used for synthetic-shift supervision.

These outputs are intermediate forensic observations rather than final manipulation probabilities or intervals. Boundary-AEP does not contain a learned reliability head. Extraction-quality metadata, optimization reliability, and BLA-AMR or Safe-AMR controller states are not part of the inference-time evidence contract.

B.3 Supervision and Boundary-AEP Objective

For modality $m \in \{a, v\}$, define the anchor abnormality target

$$y_k^m = \begin{cases} 1, & \exists\, g \in \mathcal{G}^m \text{ such that } |I_k \cap g| > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{35}$$

where $\mathcal{G}^m$ is the ground-truth interval set for modality m, and $|I_k \cap g|$ denotes the temporal duration of the intersection between anchor I_k and interval g.

We define $y_0^m = y_{K+1}^m = 0$. The undilated onset and offset targets are

$$y_{k,\text{start}}^m = y_k^m \left(1 - y_{k-1}^m\right), \qquad y_{k,\text{end}}^m = y_k^m \left(1 - y_{k+1}^m\right). \tag{36}$$

To reduce anchor-discretization artifacts, a one-anchor tolerance is applied:

$$\bar{y}_{k,\text{start}}^m = \max_{\substack{1\le j\le K\\ |j-k|\le 1}} y_{j,\text{start}}^m, \qquad \bar{y}_{k,\text{end}}^m = \max_{\substack{1\le j\le K\\ |j-k|\le 1}} y_{j,\text{end}}^m. \tag{37}$$

The tolerance is defined in anchor units. Its duration therefore follows the adaptive anchor width for clips requiring $K_{\max}$ windows.

The interval-support target is

$$q_k^m = \begin{cases} \max\limits_{\substack{j:\, y_j^m = 1\\ |k-j|\le 1}} (1 - 0.25|k-j|), & \mathcal{J}_k^m \neq \emptyset, \\ 0, & \mathcal{J}_k^m = \emptyset, \end{cases} \tag{38}$$

where $\mathcal{J}_k^m = \{j \mid y_j^m = 1,\ |k-j| \le 1\}$ denotes the set of positive neighboring anchors. Thus, an annotated positive anchor receives target 1, while an immediately adjacent anchor receives target 0.75.

For cross-modal supervision, define the modality-specific boundary indicator

$$y_{k,\text{bnd}}^m = \max\left(y_{k,\text{start}}^m, y_{k,\text{end}}^m\right). \tag{39}$$

The local consistency-violation target is

$$y_{k,\text{cons}} = \max\left(y_k^a \oplus y_k^v, y_{k,\text{bnd}}^a \oplus y_{k,\text{bnd}}^v\right), \tag{40}$$

where $\oplus$ denotes exclusive OR. Despite the subscript "cons," $y_{k,\text{cons}} = 1$ denotes a local audio–visual inconsistency. The lag target $\tilde{\ell}_k \in \mathbb{R}$ is generated from the known synthetic temporal shift applied during lag supervision; an unshifted example has $\tilde{\ell}_k = 0$.

The grouped Boundary-AEP objective follows the same definition as the main paper:

$$\mathcal{L}_{\text{AEP}} = \mathcal{L}_{\text{abn}} + \mathcal{L}_{\text{bnd}} + \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{recon}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{lag}}\mathcal{L}_{\text{lag}}, \tag{41}$$

where:

- $\mathcal{L}_{\text{abn}}$ is the aggregate modality-specific loss for $p_{k,\text{abn}}^m$ against y_k^m;
- $\mathcal{L}_{\text{bnd}}$ is the aggregate onset/offset loss for $p_{k,\text{start}}^m$ and $p_{k,\text{end}}^m$ against $\bar{y}_{k,\text{start}}^m$ and $\bar{y}_{k,\text{end}}^m$;
- $\mathcal{L}_{\text{int}}$ is the aggregate modality-specific interval-support loss for $p_{k,\text{int}}^m$ against q_k^m;
- $\mathcal{L}_{\text{recon}}$ is the temporal reconstruction objective associated with F_{rec}^m and $\varepsilon_{k,\text{rec}}^m$;
- $\mathcal{L}_{\text{cons}}$ supervises $p_{k,\text{cons}}$ using $y_{k,\text{cons}}$;
- $\mathcal{L}_{\text{lag}}$ supervises the estimated lag ℓ_k using the synthetic target $\tilde{\ell}_k$.

The coefficients λ_{cons} and λ_{lag} balance the two cross-modal auxiliary losses. Padded anchors are excluded from loss reduction. For a per-anchor term ℓ_k and a validity mask M_k, the reduction is

$$\text{MaskedMean}(\ell, M) = \frac{\sum_k M_k \ell_k}{\max(1, \sum_k M_k)}.$$

This reduction is applied separately to each supervised field before combining the task losses. A missing auxiliary target does not become a negative example: its contribution is masked out. The reconstruction residual and the lag estimate are intermediate evidence signals, not additional manipulation labels.

BLA-AMR may apply bounded weights to modality-specific training losses and sample selection. It does not modify the evidence fields in Eqs. (33) and (34), nor does it alter the inference-time procedure. Its complete training-only formulation is described in Appendix D. Computational and serialization conventions are summarized in Appendix F.

B.4 Deterministic Evidence Pack Construction

An Evidence Card is the structured record associated with one temporal anchor:

$$c_k = (\mathrm{id}_k, I_k, \mathbf{e}_k^a, \mathbf{e}_k^v, \mathbf{e}_k^{av}, \boldsymbol{\eta}_k)\,, \tag{42}$$

where id_k is the persistent identifier of anchor I_k, and $\boldsymbol{\eta}_k$ contains non-learned availability and missingness metadata required to interpret absent or invalid media observations. The metadata $\boldsymbol{\eta}_k$ is not a learned forensic prediction and is not treated as manipulation evidence.

The complete Evidence Pack is the ordered sequence

$$\mathcal{E} = \langle c_1, c_2, \ldots, c_K \rangle\,. \tag{43}$$

The order of the cards follows the temporal order of the anchors and is preserved in all downstream serializations.

All probability-like Boundary-AEP scores retain their original $[0, 1]$ values. The lag variable ℓ_k retains its signed numeric representation and is not mapped to $[0, 1]$. When categorical verbalization is enabled, a fixed parameter-free binning rule maps probability-like scores to qualitative descriptions such as `weak`, `moderate`, or `strong`. The readable renderer uses five fixed bins: `negligible` for $[0, 0.05)$, `weak` for $[0.05, 0.25)$, `moderate` for $[0.25, 0.50)$, `strong` for $[0.50, 0.75)$, and `very strong` for $[0.75, 1]$. These bins describe score magnitude; they are not decision thresholds or calibrated probabilities of manipulation. Missing values are marked unavailable rather than assigned to the lowest bin. These descriptions supplement rather than replace the original numeric values.

Two deterministic modality-focused views are generated from the same Evidence Pack:

$$\mathcal{E}^a = \rho_a\,(\mathcal{E})\,, \qquad \mathcal{E}^v = \rho_v\,(\mathcal{E})\,, \tag{44}$$

where ρ_a and ρ_v are parameter-free rendering functions for the audio and visual localization branches, respectively. The audio-focused view prioritizes audio abnormality, audio boundary, and audio interval-support fields, while the visual-focused view prioritizes the corresponding visual fields. Both views preserve the complete ordered anchor timeline and may retain cross-modal consistency-violation and lag information as context. Cross-modal evidence alone cannot assign the responsible manipulation modality.

The Evidence Pack and its rendered views never contain ground-truth modality states, ground-truth intervals, target labels, final localization answers, or BLA-AMR/Safe-AMR controller states. Non-learned extraction-quality information may appear only as availability or missingness metadata and is never presented as learned forensic evidence. At inference time, the complete pack is generated solely from the input audio A, input video V, and frozen feature-extraction and Boundary-AEP components. Therefore, the same frozen Evidence Pack can be supplied to both localization branches without target-label leakage.

B.5 Illustrative Evidence Pack and Modality-Focused Rendering

We provide an illustrative example to clarify the distinction between learned Boundary-AEP evidence and deterministic Evidence Pack metadata. The values below are used only to demonstrate the schema and do not correspond to ground-truth annotations or reported benchmark results. Only two consecutive Evidence Cards are shown for readability; the actual localization input preserves the complete ordered sequence of K anchors.

In the following example, the fields under `audio`, `visual`, and `audio_visual` are learned Boundary-AEP outputs. The schema version, clip duration, anchor identifiers, temporal ranges, media references, and missingness indicators are deterministic metadata added during Evidence Pack construction.

```
{
  "schema_version": "evidence_pack_v1",
  "clip_id": "illustrative_clip",
  "duration": 6.0,
  "anchors": [
    {
```

```
      "anchor_id": "A006",
      "time": {
        "start": 3.0,
        "end": 3.5
      },
      "audio": {
        "abnormality": 0.71,
        "start_boundary": 0.68,
        "end_boundary": 0.11,
        "interval_support": 0.73
      },
      "visual": {
        "abnormality": 0.09,
        "start_boundary": 0.06,
        "end_boundary": 0.08,
        "interval_support": 0.07
      },
      "audio_visual": {
        "consistency_disagreement": 0.24,
        "signed_lag_s": 0.0
      },
      "media_reference": {
        "spectrogram_region": "SP_A006",
        "center_keyframe": "KF_A006_C"
      },
      "missing_fields": []
    },
    {
      "anchor_id": "A007",
      "time": {
        "start": 3.5,
        "end": 4.0
      },
      "audio": {
        "abnormality": 0.84,
        "start_boundary": 0.21,
        "end_boundary": 0.76,
        "interval_support": 0.82
      },
      "visual": {
        "abnormality": 0.10,
        "start_boundary": 0.07,
        "end_boundary": 0.05,
        "interval_support": 0.08
      },
      "audio_visual": {
        "consistency_disagreement": 0.29,
        "signed_lag_s": 0.0
      },
      "media_reference": {
        "spectrogram_region": "SP_A007",
        "center_keyframe": "KF_A007_C"
      },
      "missing_fields": []
    }
  ]
}
```


The example contains an audio-oriented evidence pattern. Anchor `A006` has an elevated audio start-boundary response, while anchors `A006` and `A007` both have elevated audio abnormality and interval-support responses. Anchor `A007` has a stronger audio end-boundary response. In contrast, the visual responses over the same anchors remain low. These observations identify temporal locations that merit closer audio inspection, but they do not establish that the clip is manipulated and do not constitute a final predicted interval.

The same Evidence Pack is deterministically rendered into an audio-focused view before being provided to the audio localization branch. The following listing shows the corresponding excerpt; all other anchors remain present in the complete input.

```
BOUNDARY EVIDENCE PACK
View: AUDIO-FOCUSED
Clip duration: 6.00 seconds
Displayed excerpt: A006-A007
Complete timeline: retained in the actual input

Interpretation rules:
- Evidence is an inspection guide, not a final detector output.
- Probability-like Boundary-AEP scores lie in [0,1].
- Signed lag values retain their signed temporal units.
- Boundary responses indicate proximity to a possible onset or offset.
- Interval support indicates temporal continuity of a candidate pattern.
- Anchor ranges are reference windows, not fixed output boundaries.
- The final decision must be verified against the original audio and
  complete audio-visual context.

Anchor A006 [3.00, 3.50]
- Audio abnormality: 0.71
- Audio start-boundary response: 0.68
- Audio end-boundary response: 0.11
- Audio interval support: 0.73
- Visual abnormality: 0.09
- Audio-visual consistency disagreement: 0.24
- Signed audio-visual lag: 0.00 seconds
- Available audio reference: SP_A006

Anchor A007 [3.50, 4.00]
- Audio abnormality: 0.84
- Audio start-boundary response: 0.21
- Audio end-boundary response: 0.76
- Audio interval support: 0.82
- Visual abnormality: 0.10
- Audio-visual consistency disagreement: 0.29
- Signed audio-visual lag: 0.00 seconds
- Available audio reference: SP_A007

Deterministic inspection note:
A006 and A007 exhibit elevated audio abnormality and interval support.
The evidence is more start-oriented at A006 and more end-oriented at
A007, while visual abnormality remains low. Inspect the original audio
around these anchors before producing any manipulation interval.
Do not copy anchor boundaries directly as final timestamps.
```

This rendering does not add a fake label, a responsible-modality decision, or a final temporal interval. The audio localizer remains responsible for inspecting the original media and producing continuous timestamps under the structured output protocol.

The corresponding visual-focused excerpt retains the same anchors but prioritizes the visual fields:

```
VISUAL BOUNDARY EVIDENCE PACK (schema illustration)
Displayed excerpt: A006-A007; full clip duration: 6.00 s
A006 [3.00, 3.50]: visual abnormality 0.09;
  start 0.06; end 0.08; interval support 0.07.
A007 [3.50, 4.00]: visual abnormality 0.10;
  start 0.07; end 0.05; interval support 0.08.
Inspect the original frames. Audio responses and synchronization
context cannot establish visual manipulation. These scores alone
do not establish either a final real or a final fake decision.
```

Table 4: Recorded evidence supplied to the two localization branches. Abn., Start, End, and Int. denote within-anchor abnormality, onset response, offset response, and interval support.

Anchor	Time (s)	Modality	Abn.	Start	End	Int.
A007	3.50–4.00	Audio	0.988	0.980	0.969	1.000
A007	3.50–4.00	Visual	0.962	0.964	0.936	1.000
A008	4.00–4.50	Audio	0.978	0.932	0.986	1.000
A008	4.00–4.50	Visual	0.957	0.912	0.959	1.000

B.6 Recorded Audio–Visual Input Example

To complement the schema illustration, Table 4 transcribes numeric fields from archived audio- and visual-branch input messages for one LAV-DF development clip (duration 6.272 seconds). Both messages contain a complete 13-anchor timeline; the table shows two anchors with strong responses in both modalities. Values are copied at the precision present in those messages. This is an input-side example, not an additional evaluation result or a decoded localization output. In particular, the illustrative responses in Appendix C are not presented as predictions recovered from these messages.

The archived messages report audio–visual mismatch scores of 0.001 and 0.007, respectively, and an estimated lag of zero for both anchors. Low mismatch does not imply authentic content: the two modalities can have strong local anomaly responses without a large synchronization discrepancy. The temporal ranges in the table identify inspection windows, not exact forgery endpoints.

C Evidence-Grounded MLLM Localization and Diagnosis

This section describes how EviPatrol combines the original media with the Evidence Pack to produce modality-specific manipulation intervals, a deterministic temporal union, and evidence-grounded diagnostic reports. Localization and diagnostic generation are separated: localization outputs are first parsed and frozen, after which the diagnostic module may describe, but cannot modify, the localization result.

C.1 Split-Modal Input and LoRA Localization

Let A_i, V_i, and $\mathcal{E}_i$ denote the original audio, original video, and ordered Evidence Pack of sample i, respectively. For branch $m \in \{a, v\}$, where a and v denote audio and visual localization responsibilities, the conditional input is

$$X_i^m = (\Pi_m, A_i, V_i, \rho_m(\mathcal{E}_i)), \tag{45}$$

where Π_m is the modality-responsibility instruction and $\rho_m \in \{\rho_a, \rho_v\}$ is the corresponding parameter-free Evidence Pack renderer defined in Eq. (44).

Both localization branches receive the same complete audio–visual clip. They differ only in their Evidence Pack rendering, trainable LoRA parameters, supervision targets, and output responsibilities. An illustrative input assembly for a 6.272-second clip is shown below.

```
Audio localization branch:
  1. Original 6.272-second audio-video clip
  2. Audio-focused Boundary-AEP evidence view
  3. Audio responsibility prompt and output schema

Visual localization branch:
  1. The same original audio-video clip
  2. Visual-focused Boundary-AEP evidence view
  3. Visual responsibility prompt and output schema
```

Let Θ denote the frozen Qwen2.5-Omni-7B backbone parameters, and let Δ_m denote the independently trained LoRA parameters for branch m. The structured branch output is

$$\widehat{Y}_i^m = F_{\Theta,\Delta_m}(X_i^m), \tag{46}$$

where F_{Θ,Δ_m} denotes the frozen backbone equipped with the modality-specific LoRA adapter.

For supervised adaptation, let

$$Y_i^m = \left(Y_{i,1}^m, Y_{i,2}^m, \ldots, Y_{i,L_{Y,i}^m}^m\right) \tag{47}$$

denote the structured target-token sequence, where $L_{Y,i}^m$ is its token length and $Y_{i,t}^m$ is the token at position t. Let $\mathcal{J}_i^m \subseteq \{1, \ldots, L_{Y,i}^m\}$ denote the assistant-token positions included in the training loss. The modality-specific SFT objective is

$$\mathcal{L}_{\mathrm{SFT}}^m = -\sum_{t \in \mathcal{J}_i^m} \log p_{\Theta,\Delta_m}\left(Y_{i,t}^m \mid X_i^m, Y_{i,<t}^m\right), \tag{48}$$

where $Y_{i,<t}^m$ is the target prefix before position t, and p_{Θ,Δ_m} is the next-token distribution produced by the adapted model.

The audio branch is permitted to emit intervals only in the audio interval list, while the visual branch is permitted to emit intervals only in the visual interval list. The non-responsible interval list must remain empty. Information from the other modality may still provide temporal and synchronization context because both branches receive the complete audio–visual input. The output responsibility is enforced by the prompt contract, modality-specific supervision, deterministic parser, and the training-time shortcut mitigation described in Appendix D. This constraint limits the permitted output of each branch; it does not assume that cross-modal contextual influence is causally absent.

C.2 Localization Prompts and Output Contract

The localization prompts instruct each branch to inspect the original media, treat Boundary-AEP evidence as an inspection guide, and return a compact localization-only JSON object. Diagnostic rationales are not generated at this stage. They are produced only after localization has been frozen.

The following listings show illustrative prompt instances for a 6.272-second clip. At inference time, the clip duration and modality-focused Evidence Pack view are inserted deterministically. The `INSERT` markers below are template slots, not text sent to the model: each is replaced by the corresponding serialized evidence view. Both branches receive the same original clip and duration, with branch-specific instructions and evidence ordering. A schematic message assembly is given in Appendix F.2.

```
Analyze this clip as the AUDIO localization branch for deepfake
forensics.

Clip duration: 6.272 seconds.

The complete audio-video input is provided for temporal and
synchronization context. This branch is responsible only for audio
authenticity and audio manipulation localization.

The Boundary-AEP evidence block defines each score before listing
anchor observations. Interpret within-anchor abnormality as local
forensic evidence, start/end responses as boundary-proximity evidence,
and interval support only as an inspection cue.

Inspect the original audio before making the audio decision and
producing continuous audio timestamps. Visual and audio-visual context
may help with temporal alignment, but cannot alone establish an
audio-fake decision.

Do not report visual manipulation, and keep the visual interval list
empty. Do not copy visual artifacts or transcript events into the audio
interval list. Do not generate diagnostic prose in this stage.

<INSERT AUDIO-FOCUSED BOUNDARY EVIDENCE VIEW>

Return one compact valid JSON object only:
```

```
{
  "s": "RR | FR",
  "c": 0.0,
  "a": [{
    "start": 0.0,
    "end": 0.0,
    "confidence": 0.0,
    "evidence": ["supporting_anchor_id"]
  }],
  "v": [],
  "e": ["optional_audio_cue_tag"]
}

If the audio is real, return "s":"RR" and "a":[].
If the audio is fake, return "s":"FR".
Each predicted audio interval must cite one or more existing audio
evidence anchors whose time ranges overlap the interval.
Return JSON only.



Analyze this clip as the VISUAL localization branch for deepfake
forensics.

Clip duration: 6.272 seconds.

The complete audio-video input is provided for temporal and
synchronization context. This branch is responsible only for visible
video manipulation and visual localization.

Interpret within-anchor abnormality as local forensic evidence,
start/end responses as boundary-proximity evidence, and interval
support only as an inspection cue.

Inspect the original frames before making the visual decision and
producing continuous visual timestamps. Audio and audio-visual context
may help with temporal alignment, but cannot alone establish a
visual-fake decision.

Do not report audio manipulation, and keep the audio interval list
empty. Do not copy audio segments, transcript events, or acoustic
artifacts into the visual interval list. Emit a visual interval only
when the original frames provide visual support. Do not generate
diagnostic prose in this stage.

<INSERT VISUAL-FOCUSED BOUNDARY EVIDENCE VIEW>

Return one compact valid JSON object only:
{
  "s": "RR | RF",
  "c": 0.0,
  "a": [],
  "v": [{
    "start": 0.0,
    "end": 0.0,
    "confidence": 0.0,
    "evidence": ["supporting_anchor_id"]
  }],
  "e": ["optional_visual_cue_tag"]
}

If the visual stream is real, return "s":"RR" and "v":[].
If the visual stream is fake, return "s":"RF".
Each predicted visual interval must cite one or more existing visual
```

```
evidence anchors whose time ranges overlap the interval.
Return JSON only.
```

Field `s` is branch-local generated metadata. In the audio branch, the visual position is fixed to `R`; in the visual branch, the audio position is fixed to `R`. It is not used to determine the evaluated global state. The final RR/RF/FR/FF state is reconstructed exclusively from valid audio and visual interval presence under the segments-only protocol.

Field `c` stores optional branch-level confidence metadata and is not used to rank temporal proposals. Temporal AP uses the interval-level `confidence` field. Field `e` stores optional model-generated cue tags for output inspection; these tags do not replace Boundary-AEP evidence fields, do not affect localization metrics, and are not treated as deterministic evidence provenance.

C.3 Deterministic Parser and Illustrative Responses

The parser is fixed before evaluation. It first extracts the accepted structured object and checks branch responsibility. Audio outputs must contain an empty visual interval list, while visual outputs must contain an empty audio interval list. Accepted interval entries are converted to the raw prediction set $\widehat{\mathcal{S}}_i^m$ defined in Eq. (15).

The deterministic validation and sanitization operator in Eq. (16) then produces the valid interval set $\overline{\mathcal{S}}_i^m$. Endpoint ordering, temporal clipping, finite-value checks, duration validation, confidence handling, and complete parse failures follow the common protocol in Appendix A. The evaluated modality state is reconstructed from $\overline{\mathcal{S}}_i^m$ using Eq. (17); generated field `s` cannot override this decision.

For later notation, write the valid interval set as

$$\overline{\mathcal{S}}_i^m = \left\{ \left(\bar{b}_{ij}^m, \bar{e}_{ij}^m, \bar{c}_{ij}^m, \overline{\mathcal{K}}_{ij}^m \right) \right\}_{j=1}^{\bar{N}_i^m}, \tag{49}$$

where $\bar{N}_i^m$ is the number of valid intervals, $\bar{b}_{ij}^m$ and $\bar{e}_{ij}^m$ are the sanitized endpoints, $\bar{c}_{ij}^m \in [0,1]$ is the sanitized interval confidence, and $\overline{\mathcal{K}}_{ij}^m$ is the retained set of syntactically valid Evidence Card anchor references.

Illustrative modality-specific responses. The following JSON objects are schema-consistent illustrations of the audio and visual branch outputs. They are not actual model predictions for the Evidence Pack example in Appendix B, do not correspond to a benchmark sample, and are not used in any reported metric.

```
{
  "s": "FR",
  "c": 0.93,
  "a": [{
    "start": 3.500,
    "end": 4.224,
    "confidence": 0.91,
    "evidence": ["A007", "A008"]
  }],
  "v": [],
  "e": [
    "spectral_discontinuity",
    "speaker_characteristic_shift"
  ]
}

{
  "s": "RF",
  "c": 0.92,
```

```
  "a": [],
  "v": [{
    "start": 3.500,
    "end": 4.224,
    "confidence": 0.90,
    "evidence": ["A007", "A008"]
  }],
  "e": [
    "mouth_motion_anomaly",
    "face_boundary_artifact"
  ]
}
```

In both illustrations, the cited anchors overlap the continuous predicted interval. The cue tags in field `e` are illustrative branch outputs rather than Boundary-AEP field names or verified diagnostic conclusions. Free-form text outside the accepted JSON object cannot create, replace, or modify a localization interval.

C.4 Deterministic Union and Confidence Handling

The deterministic union is constructed from the valid audio and visual sets $\overline{\mathcal{S}}_i^a$ and $\overline{\mathcal{S}}_i^v$. Define the source-index set

$$\mathcal{P}_i = \left\{(m, j) \,\middle|\, m \in \{a, v\},\ 1 \le j \le \bar{N}_i^m\right\}, \tag{50}$$

where each pair (m, j) identifies the j-th valid interval of modality m.

Construct an undirected temporal-connectivity graph over $\mathcal{P}_i$. Two distinct source indices are connected when their intervals overlap or exactly touch:

$$(m, j) \leftrightarrow (m', j') \quad \Longleftrightarrow \quad \bar{b}_{ij}^m \le \bar{e}_{ij'}^{m'} \ \wedge\ \bar{b}_{ij'}^{m'} \le \bar{e}_{ij}^m, \tag{51}$$

where $m, m' \in \{a, v\}$. Let $\mathcal{C}_{ir} \subseteq \mathcal{P}_i$ denote the r-th connected component of this graph, where $r \in \{1, \dots, J_i^u\}$ and J_i^u is the number of resulting union components.

For component $\mathcal{C}_{ir}$, the union endpoints are

$$\hat{b}_{ir}^u = \min_{(m,j)\in\mathcal{C}_{ir}} \bar{b}_{ij}^m, \qquad \hat{e}_{ir}^u = \max_{(m,j)\in\mathcal{C}_{ir}} \bar{e}_{ij}^m, \tag{52}$$

and its direct confidence is

$$\hat{c}_{ir}^{u,\mathrm{dir}} = \max_{(m,j)\in\mathcal{C}_{ir}} \bar{c}_{ij}^m. \tag{53}$$

The union provenance and retained anchor-reference set are

$$\Omega_{ir}^u = \mathcal{C}_{ir}, \qquad \widehat{\mathcal{K}}_{ir}^u = \bigcup_{(m,j)\in\Omega_{ir}^u} \overline{\mathcal{K}}_{ij}^m, \tag{54}$$

where Ω_{ir}^u records all contributing modality-specific intervals. The deterministic union is

$$\widehat{\mathcal{U}}_i = \left\{ \left(\hat{b}_{ir}^u, \hat{e}_{ir}^u, \hat{c}_{ir}^{u,\mathrm{dir}}, \widehat{\mathcal{K}}_{ir}^u, \Omega_{ir}^u \right) \right\}_{r=1}^{J_i^u}. \tag{55}$$

Union construction does not alter any source modality-specific interval and cannot create a union interval without at least one valid audio or visual source interval. It forms temporal envelopes over connected source support and is not a separately learned third localizer.

When a learned confidence-only scorer is selected under the frozen protocol in Appendix A, it may replace $\hat{c}_{ir}^{u,\mathrm{dir}}$ with a calibrated confidence value. The scorer cannot modify the union endpoints, evidence references, provenance, interval count, or modality states. If the scorer is not selected, the direct maximum-source confidence is retained.

Illustrative deterministic fusion. Applying the deterministic union rule to the two illustrative modality-specific responses above yields the following representation:

```
{
  "video_state": "FF",
  "audio_fake_segments": [[3.500, 4.224]],
  "visual_fake_segments": [[3.500, 4.224]],
  "union_fake_segments": [[3.500, 4.224]]
}
```

This object is a deterministic composition of the two branch outputs, not a generated response from a third model. The global state `FF` follows from the presence of at least one valid audio interval and at least one valid visual interval.

C.5 Localization-Preserving Diagnosis

Diagnostic generation begins only after both localization branches, the deterministic parser, confidence handling, modality states, and union geometry have been frozen.

Let

$$B_i = \left(\hat{z}_i^a, \hat{z}_i^v, \overline{\mathcal{S}}_i^a, \overline{\mathcal{S}}_i^v, \widehat{\mathcal{U}}_i\right) \tag{56}$$

denote the immutable localization record for sample i.

For the j-th valid interval of modality m, define the verified citation set

$$\mathcal{A}_{ij}^m = \left\{ \mathrm{id}_k \;\middle|\; \mathrm{id}_k \in \overline{\mathcal{K}}_{ij}^m, \quad \left|I_k \cap [\bar{b}_{ij}^m, \bar{e}_{ij}^m]\right| > 0 \right\}, \tag{57}$$

where $\mathcal{A}_{ij}^m$ contains only cited Evidence Card anchors that both exist in $\mathcal{E}_i$ and temporally overlap the frozen interval. Let $\mathcal{F}_{ij}^m$ denote the set of available modality-compatible Boundary-AEP fields retrieved from those cards.

The deterministic provenance record is

$$\boldsymbol{\pi}_{ij}^m = \left(m, [\bar{b}_{ij}^m, \bar{e}_{ij}^m], \mathcal{A}_{ij}^m, \mathcal{F}_{ij}^m\right). \tag{58}$$

Let Δ_{exp} denote the explanation-adapter parameters. The frozen Qwen2.5-Omni backbone equipped with this adapter is denoted by $G_{\Theta,\Delta_{\mathrm{exp}}}$. It conditions on the original media, Evidence Pack, immutable localization record, and deterministic provenance:

$$\mathbf{P}_i = G_{\Theta,\Delta_{\mathrm{exp}}}\left(A_i, V_i, \mathcal{E}_i, B_i, \left\{\boldsymbol{\pi}_{ij}^m\right\}_{m,j}\right), \tag{59}$$

where $\mathbf{P}_i$ contains only generated diagnostic suffixes. For interval (m, j), let $\mathbf{p}_{ij}^m$ denote its generated textual fields, such as observation, assessment, alternative, and conclusion. The final interval-level rationale is formed by deterministic assembly:

$$\mathbf{r}_{ij}^m = \mathrm{Assemble}\left(\boldsymbol{\pi}_{ij}^m, \mathbf{p}_{ij}^m\right). \tag{60}$$

The explanation adapter does not generate or modify interval endpoints, modality labels, state values, anchor identifiers, evidence-field names, confidence values, or union geometry.

Illustrative explanation prompt. The following example uses the frozen deterministic fusion shown above. The localization and provenance blocks are inserted by the system and are immutable.

```
Generate an evidence-grounded forensic explanation for the frozen
localization result below.

Frozen localization:
- State: FF
- Audio fake interval: 3.500-4.224 seconds
- Visual fake interval: 3.500-4.224 seconds
- Union interval: 3.500-4.224 seconds

Frozen audio provenance:
```

```
- Anchors: A007, A008
- Evidence fields: audio_abnormality, audio_start_boundary,
  audio_end_boundary, audio_interval_support

Frozen visual provenance:
- Anchors: A007, A008
- Evidence fields: visual_abnormality, visual_start_boundary,
  visual_end_boundary, visual_interval_support

The localization, modality decisions, provenance, and evidence-field
names are immutable. Do not add, remove, expand, shrink, relabel, or
rename them.

For each modality:
1. describe a concrete observation from the corresponding raw modality;
2. explain how its temporal pattern supports the frozen decision;
3. consider a plausible natural alternative;
4. explain why the manipulation interpretation is better supported.

Audio explanations may use only raw audio and audio evidence.
Visual explanations may use only raw frames and visual evidence.
Audio-visual synchronization may provide temporal context but cannot
assign modality ownership.

Do not justify a conclusion merely by repeating numerical scores.
If the available evidence is insufficient, abstain instead of
inventing forensic details. Return valid JSON only.
```

Illustrative assembled diagnostic report. The following object illustrates the report after deterministic provenance assembly and before verifier release. It is a schema example, not an actual decoded result for a benchmark sample. It shows the assembled report rather than an object generated in full by the explanation adapter. The adapter supplies only `observation`, `assessment`, `alternative`, and `why` for each fixed interval; the deterministic assembler owns the interval, modality, decision, anchor, and evidence-field keys.


```
{
  "frozen_localization": {
    "state": "FF",
    "audio": [[3.500, 4.224]],
    "visual": [[3.500, 4.224]],
    "union": [[3.500, 4.224]]
  },
  "diagnosis": [
    {
      "span": [3.500, 4.224],
      "modality": "audio",
      "decision": "fake",
      "anchors": ["A007", "A008"],
      "evidence_fields": [
        "audio_abnormality",
        "audio_start_boundary",
        "audio_end_boundary",
        "audio_interval_support"
      ],
      "pattern": "bounded_sustained_region",
      "observation": "Within the frozen interval, the raw audio contains
    an abrupt acoustic transition followed by a sustained change in
    spectral texture and speaker characteristics relative to the adjacent
     context.",
      "assessment": "The cited anchors contain sustained audio
    abnormality and interval support together with onset- and offset-
    oriented boundary responses, forming a temporally coherent pattern
    inside the frozen interval.",
```

```
      "alternative": "Natural prosody, recording changes, or compression
    may create a brief acoustic discontinuity, but they are less
    consistent with a sustained localized change bounded by two temporal
    transitions.",
      "why": "The acoustic change persists inside the frozen interval and
     returns toward the surrounding pattern after its boundary."
    },
    {
      "span": [3.500, 4.224],
      "modality": "visual",
      "decision": "fake",
      "anchors": ["A007", "A008"],
      "evidence_fields": [
        "visual_abnormality",
        "visual_start_boundary",
        "visual_end_boundary",
        "visual_interval_support"
      ],
      "pattern": "bounded_sustained_region",
      "observation": "Within the frozen interval, mouth motion and the
    facial boundary become temporally unstable relative to the
    surrounding frames.",
      "assessment": "The cited anchors contain sustained visual
    abnormality and interval support together with boundary-oriented
    responses, supporting a localized visible manipulation.",
      "alternative": "Head motion, expression changes, or compression may
     cause brief visual artifacts, but they are less consistent with a
    sustained anomaly confined to the frozen interval.",
      "why": "The visible irregularities persist inside the frozen
    interval and diminish in the surrounding frames."
    }
  ],
  "summary": "Both modalities contain a localized manipulation between
    3.500 and 4.224 seconds, with modality-specific evidence in the
    original audio and frames.",
  "explanation_status": "pending_verification"
}
```


In the assembled report, the fields `span`, `modality`, `decision`, `anchors`, and `evidence_fields` are inserted deterministically from the frozen localization and provenance. Only the natural-language diagnostic fields are generated by the explanation adapter.

Let

$$\mathcal{C}_{\text{ver}} = \{\text{exist}, \text{overlap}, \text{modality}, \text{decision}\} \tag{61}$$

denote the fixed verifier checks for anchor existence, temporal overlap, modality compatibility, and agreement with the frozen localization decision. For check $c \in \mathcal{C}_{\text{ver}}$, let $\gamma_{ij,c}^{m} \in \{0, 1\}$ indicate whether the check passes for interval (m, j). The interval-level release gate is

$$\Gamma_{ij}^{m} = \prod_{c \in \mathcal{C}_{\text{ver}}} \gamma_{ij,c}^{m}. \tag{62}$$

If $\Gamma_{ij}^{m} = 1$, the corresponding rationale is released. If $\Gamma_{ij}^{m} = 0$, its detailed prose is withheld and an explanation abstention is recorded. Verified interval-level rationales are aggregated into the video-level diagnostic report R_i.

For example, a report-level abstention can be serialized as follows. This is an output-contract illustration, not an observed failure case; `failed_checks` is populated by the verifier, not the language model. An unavailable explanation never becomes a negative localization prediction.


```
{
  "frozen_localization": {
    "state": "FF",
    "audio": [[3.500, 4.224]],
    "visual": [[3.500, 4.224]],
```

```
    "union": [[3.500, 4.224]]
  },
  "diagnosis": [],
  "summary": null,
  "explanation_status": "abstained",
  "failed_checks": ["exist"]
}
```


The deterministic verifier checks structural and provenance consistency; it does not convert generated natural-language observations into a causal guarantee. Regardless of the verifier outcome, the diagnostic stage cannot modify $\overline{\mathcal{S}}_i^a$, $\overline{\mathcal{S}}_i^v$, $\hat{z}_i^a$, $\hat{z}_i^v$, or $\widehat{\mathcal{U}}_i$. Explanation quality and abstention therefore have no effect on temporal localization AP.

C.6 End-to-End Inference Algorithm

Algorithm 1 Evidence-grounded split-modal inference

Require: Audio A_i, video V_i, and duration T_i
Ensure: Valid modality-specific intervals, deterministic union, and optional diagnostic report

Construct the synchronized anchor sequence $\mathcal{I}_i$
Extract frozen MARLIN and Whisper representations
Apply the frozen shift-alignment adapter
Compute Boundary-AEP evidence for all valid anchors
Construct the ordered Evidence Pack $\mathcal{E}_i$
for $m \in \{a, v\}$ **do**
    Render the modality-focused view $\rho_m(\mathcal{E}_i)$
    Construct X_i^m using Eq. (45)
    Generate $\widehat{Y}_i^m \leftarrow F_{\Theta,\Delta_m}(X_i^m)$
    Parse and sanitize $\widehat{Y}_i^m$ to obtain $\overline{\mathcal{S}}_i^m$
    Reconstruct $\hat{z}_i^m$ using the segments-only rule
end for
Reconstruct the global $\mathrm{RR}/\mathrm{RF}/\mathrm{FR}/\mathrm{FF}$ state
Construct $\widehat{\mathcal{U}}_i$ and provenance using Eqs. (52)–(55)
if a confidence-only scorer is selected **then**
    Update union confidences without changing union geometry
end if
if diagnostic reporting is enabled **then**
    Construct the immutable localization record B_i
    Construct deterministic interval-level provenance $\boldsymbol{\pi}_{ij}^m$
    Generate diagnostic suffixes using $G_{\Theta,\Delta_{\mathrm{exp}}}$
    Assemble and verify interval-level rationales
    Aggregate verified rationales into R_i, abstaining on failed intervals
end if
return $\overline{\mathcal{S}}_i^a, \overline{\mathcal{S}}_i^v, \widehat{\mathcal{U}}_i$, and optional R_i

D Modality Bias Mitigation

Both BLA-AMR and Safe-AMR are training-only mechanisms. BLA-AMR balances modality-specific evidence learning in Boundary-AEP, while Safe-AMR reduces undesirable cross-modal shortcuts during visual MLLM adaptation. Neither mechanism introduces additional inference-time branches or computational overhead.

D.1 Boundary-Learnability-Aware Adaptive Modality Rebalancing

BLA-AMR is designed to alleviate asymmetric modality learning during Boundary-AEP optimization. Let $\mathcal{L}_{m,t}^{\mathrm{core}}$ denote the core evidence-learning loss of modality $m \in \{a, v\}$ at training step t,

including abnormality, boundary, and interval-support supervision. After a neutral warm-up stage, the relative learning progress is estimated as

$$P_{m,t} = \mathrm{clip}_{[0,1]}\left(1 - \mathrm{EMA}\left[\frac{\mathcal{L}^{\mathrm{core}}_{m,t}}{\mathcal{L}^{\mathrm{core}}_{m,0} + \epsilon}\right]\right), \tag{63}$$

where ϵ is a small constant for numerical stability and EMA denotes exponential moving average. A larger $P_{m,t}$ indicates faster reduction of the corresponding modality loss. The reference loss $\mathcal{L}^{\mathrm{core}}_{m,0}$ is estimated during neutral warm-up and then held fixed. All controller statistics are detached scalars. For an observed scalar x_t, $\overline{x}_t = \beta\overline{x}_{t-1} + (1-\beta)x_t$, with $0 \le \beta < 1$; the first finite observation initializes the average. Missing observations retain the previous state rather than being imputed as zero. Controls for a batch are read before its measurements update the controller.

The modality learning gap is measured by

$$L_{v,t} = [P_{a,t} - P_{v,t} - \delta]_+, \qquad L_{a,t} = [P_{v,t} - P_{a,t} - \delta]_+, \tag{64}$$

where $[x]_+ = \max(x, 0)$. In addition, gradient dominance over shared Boundary-AEP parameters is measured using the gradient norms $g_{a,t}$ and $g_{v,t}$ induced by the two modality losses:

$$\begin{aligned} D_{v,t} &= \left[\frac{g_{a,t} - g_{v,t}}{g_{a,t} + g_{v,t} + \epsilon} - \delta_g\right]_+, \\ D_{a,t} &= \left[\frac{g_{v,t} - g_{a,t}}{g_{a,t} + g_{v,t} + \epsilon} - \delta_g\right]_+. \end{aligned} \tag{65}$$

The monitored gradient norm is $g_{m,t} = \|\nabla_{\theta_{\mathrm{sh}}}\mathcal{L}^{\mathrm{core}}_{m,t}\|_2$, using the same shared fusion, temporal, and pyramid parameters for both modalities and excluding the modality-specific output heads. The smoothed dominance gate is

$$G^{\mathrm{dom}}_{m,t} = g^{\mathrm{dom}}_{\min} + (1 - g^{\mathrm{dom}}_{\min})\sigma\left(\frac{\overline{D}_{m,t} - \rho_{\mathrm{dom}}}{\tau_{\mathrm{dom}}}\right), \tag{66}$$

where $\overline{D}_{m,t} = \mathrm{EMA}(D_{m,t})$, σ is the logistic sigmoid, $\tau_{\mathrm{dom}} > 0$, and $g^{\mathrm{dom}}_{\min} \in [0, 1]$. When gradient monitoring is disabled, the gate is one.

The reference controller also moderates the boost using the recent learning-progress change $\Delta P_{m,t} = P_{m,t} - P_{m,t-w}$ over a fixed window w. Its gate is

$$G^{\mathrm{lp}}_{m,t} = g^{\mathrm{lp}}_{\min} + (1 - g^{\mathrm{lp}}_{\min})\sigma\left(\frac{\Delta P_{m,t} - \rho_{\mathrm{lp}}}{\tau_{\mathrm{lp}}}\right). \tag{67}$$

Before a complete history window is available, $\Delta P_{m,t} = 0$. The slopes, offsets, floors, and window length are fixed controller configuration values, not learnable network parameters.

The above signals are combined into bounded modality-specific reweighting factors:

$$B_{m,t} = \mathrm{clip}\left(L_{m,t}(1 - P_{m,t})G^{\mathrm{lp}}_{m,t}G^{\mathrm{dom}}_{m,t}Q_{m,t}, 0, B^{\max}_m\right), \tag{68}$$

where $Q_{a,t} = 1$. If the optional quality modifier is enabled, $Q_{v,t}$ is the clipped EMA of externally supplied visual extraction-quality metadata, aggregated over visual-positive anchors when present and otherwise over valid anchors. When this metadata is unavailable, or the modifier is disabled, $Q_{v,t} = 1$. It is not a prediction of authenticity and is not a learned reliability head. The factor is used only by the controller and is not included in the Evidence Pack. During warm-up both boosts are zero and all task multipliers are one.

For evidence task j, the loss multiplier is

$$\mu^j_{m,t} = \mathrm{clip}(1 + \kappa_j B_{m,t}, 1, \mu^{\max}_j). \tag{69}$$

The final BLA-AMR objective is

$$\mathcal{L}_{\mathrm{AEP}}^{\mathrm{BLA}} = \sum_{m\in\{a,v\}} \sum_{j} \mu_{m,t}^{j} \mathcal{L}_{m}^{j} + \mathcal{L}_{\mathrm{shared}}, \tag{70}$$

where $\mathcal{L}_{m}^{j}$ denotes the modality-specific evidence loss for task j, and $\mathcal{L}_{\mathrm{shared}}$ contains shared temporal and cross-modal regularization terms. BLA-AMR only changes training weights and does not modify the inference-time evidence schema.

D.2 SAFE-AMR

Safe-AMR is applied only during visual MLLM adaptation to reduce undesirable audio-dependent shortcuts. Under the audio–visual state order, RF and FF share the same visual-fake target, while RR and FR share the same visual-real target. Therefore, large differences between paired visual decision losses indicate possible dependence on the audio condition.

Let $d_{s,t}$ denote the EMA visual decision-prefix loss for state $s \in \{RR, RF, FR, FF\}$. The prefix is taken from the supervised assistant answer, not from the input prompt or media tokens. If $\mathcal{P}_i$ contains its first N_{pref} non-masked answer-token positions, the measured loss is

$$d_i = -\frac{1}{|\mathcal{P}_i|} \sum_{n\in\mathcal{P}_i} \log p(Y_{i,n}^{v} \mid X_i^{v}, Y_{i,<n}^{v}). \tag{71}$$

The target serialization places the visual decision before the long diagnostic suffix. Consequently, the prefix is a decision-sensitive training proxy, not a separate calibrated probability of correctness. The same prefix rule is used for the regular and counterfactual inputs. Each state's EMA is initialized by its first finite observation and updated only when that state is observed; an absent state does not receive a zero loss. The paired-state signal is inactive until all four states have been initialized. The paired-state discrepancy is

$$\Delta_t = \frac{1}{2}\left(\left|\hat{d}_{RF,t} - \hat{d}_{FF,t}\right| + \left|\hat{d}_{RR,t} - \hat{d}_{FR,t}\right|\right), \tag{72}$$

where $d_{s,0}$ is the first finite reference loss for state s, and

$$\hat{d}_{s,t} = \frac{d_{s,t}}{d_{s,0} + \epsilon}.$$

For selected training samples, Safe-AMR introduces an audio-withheld counterfactual input:

$$\widetilde{X}_i^{v} = (\emptyset^{a}, V_i, \emptyset_{\mathcal{E}}^{a}, \mathcal{E}_i^{v}), \tag{73}$$

which removes raw audio and audio-related evidence while preserving the visual stream, visual Evidence Pack, and the same visual target.

Let C_t denote the average decision sensitivity between the normal and audio-withheld branches:

$$C_t = \mathrm{EMA}\left(\frac{|d_t - \widetilde{d}_t|}{\max(|d_t|, |\widetilde{d}_t|, \epsilon)}\right), \tag{74}$$

where $\widetilde{d}_t$ is the corresponding counterfactual decision loss for the same training example. Only observed regular/counterfactual pairs update this statistic. Neither signal uses target-domain evaluation labels. The shortcut score is

$$S_t = \mathrm{clip}_{[0,1]}\left(\lambda_{\Delta}\Delta_t + \lambda_C C_t\right). \tag{75}$$

The score controls the counterfactual sampling probability:

Table 5: Four-state modality-state analysis on LAV-DF. BLA and Safe indicate whether BLA-AMR and Safe-AMR are enabled, respectively. RR, RF, FR, and FF denote real–real, real–fake, fake–real, and fake–fake audio-visual states following the audio–visual order. Each state column reports the corresponding prediction accuracy, and Avg. denotes the average accuracy over the four states.

BLA	Safe	RR	RF	FR	FF	Avg.
×	×	87.61	43.45	57.98	93.78	70.71
×	√	93.74	83.16	76.37	97.65	87.73
√	×	96.29	81.40	84.91	96.75	89.84
√	√	**98.93**	**94.17**	**92.82**	**97.87**	**95.95**

$$p_t^{\mathrm{cf}} = p_{\min} + S_t(p_{\max} - p_{\min}). \tag{76}$$

The final Safe-AMR objective is

$$\mathcal{L}_i^{\text{Safe-AMR}} = \mathcal{L}_i^{\mathrm{main}} + \mathbf{1}_i^{\mathrm{cf}} \mathcal{L}_i^{\mathrm{cf}}, \qquad \mathbf{1}_i^{\mathrm{cf}} \sim \mathrm{Bernoulli}(p_t^{\mathrm{cf}}), \tag{77}$$

where $\mathcal{L}_i^{\mathrm{main}}$ and $\mathcal{L}_i^{\mathrm{cf}}$ denote the normal and audio-withheld training objectives, respectively; any fixed counterfactual loss coefficient is included in $\mathcal{L}_i^{\mathrm{cf}}$. Both branches use the same visual supervision and answer-token mask. The Bernoulli draw is a training-time sampling operation. Gradients are not propagated through the EMA statistics, shortcut score, or sampling probability.

Safe-AMR is asymmetric by design: it is applied only to the visual adapter because the visual branch is the target of the identified audio-related shortcut. The independently trained audio adapter uses the fixed evidence-aware weighting strategy. During inference, the audio-withheld branch and all adaptive controllers are removed.

E Additional Experiments

This section provides additional analyses beyond the main experimental results. We focus on three aspects closely related to the key design choices of EviPatrol: modality bias mitigation, cross-dataset generalization, and human evaluation of diagnostic reports.

E.1 Modality Bias Mitigation Analysis

Beyond temporal localization performance, we further analyze whether BLA-AMR and Safe-AMR improve the consistency of modality-state prediction. The four manipulation states follow the audio–visual order: RR, RF, FR, and FF denote real–real, real–fake, fake–real, and fake–fake states, respectively.

BLA-AMR is designed to alleviate asymmetric modality learning during Boundary-AEP optimization, while Safe-AMR reduces undesirable cross-modal decision shortcuts during visual MLLM adaptation. To evaluate their individual contributions, we remove each component and report four-state prediction accuracy on LAV-DF.

As shown in Table 5, enabling either BLA-AMR or Safe-AMR improves the consistency of four-state prediction compared with the model without modality bias mitigation. Removing BLA-AMR mainly degrades the overall state prediction accuracy, indicating the importance of balanced evidence learning between modalities. Removing Safe-AMR causes larger degradation on paired states with identical visual targets but different audio conditions, i.e., RF/FF and RR/FR, suggesting that Safe-AMR helps reduce undesirable dependence on cross-modal conditions. The complete model achieves the best performance across all four states, demonstrating the complementary effects of the two training-time strategies.

Table 6: Cross-dataset generalization from LAV-DF to AV-Deepfake1M. Models are trained on LAV-DF and evaluated on AV-Deepfake1M without target-domain adaptation or calibration. All results are reported as AP (%).

Method	AP@0.50	AP@0.75	AP@0.95
ActionFormer	13.98	06.72	00.74
BA-TFD+	21.98	10.79	00.24
UMMAFormer	28.49	13.65	01.44
Intra-/Cross-modal Sync	25.67	16.98	00.76
EviPatrol (Ours)	**65.47**	**50.34**	**06.28**

Table 7: Human evaluation of evidence-grounded diagnosis. Und. and Trust. denote explanation understandability and trustworthiness rated on a 5-point Likert scale. Ver. Acc. denotes verification accuracy (%).

Setting	Und.	Trust.	Ver. Acc. (%)
Prediction Only	3.27	2.47	63.74
MLLM Explanation	3.81	3.42	82.91
EviPatrol Diagnosis	**4.27**	**4.07**	**87.60**

E.2 Cross-Dataset Generalization

To evaluate generalization under domain shift, we train EviPatrol on LAV-DF and directly evaluate it on AV-Deepfake1M without target-domain adaptation. All checkpoint selection, prompt choices, and threshold decisions are determined using only the LAV-DF development split.

Table 6 shows that EviPatrol maintains a clear advantage under the cross-dataset setting, especially at stricter temporal-IoU thresholds. Compared with existing localization methods, EviPatrol achieves substantially higher AP@0.50 and AP@0.75 while also improving performance under strict boundary evaluation. This suggests that explicitly structured boundary evidence provides more transferable temporal forensic cues than dataset-specific implicit representations.

E.3 Human Evaluation of Evidence-Grounded Diagnosis

To evaluate whether EviPatrol improves the interpretability and verification of deepfake localization results, we conduct a human study with 30 participants. Each participant is provided with localization results under three different settings: (1) *Prediction Only*, which presents only the predicted manipulation intervals; (2) *MLLM Explanation*, which provides a free-form explanation generated without structured evidence grounding; and (3) *EviPatrol Diagnosis*, which provides evidence-grounded diagnostic reports with modality-specific temporal evidence.

Participants evaluate each setting from two subjective perspectives: explanation understandability and trustworthiness. Both metrics are rated on a 5-point Likert scale. In addition, we introduce an objective verification task, where participants determine whether the predicted manipulation interval is supported by the provided information. The verification accuracy is calculated based on the agreement between participant decisions and the manually verified interval correctness.

As shown in Table 7, providing explanations improves both subjective understanding and objective verification compared with using localization results alone. In particular, MLLM Explanation substantially improves verification accuracy from 63.74% to 82.91%, indicating that natural-language explanations can help users interpret model decisions.

Compared with generic MLLM explanations, EviPatrol Diagnosis further improves understandability and trustworthiness by 0.46 and 0.65 points, respectively, on the 5-point Likert scale. Moreover, it achieves higher verification accuracy (87.60%), demonstrating that explicitly linking predictions

with modality-specific temporal evidence provides more reliable support for human verification than unconstrained explanations.

These results suggest that structured evidence representations not only assist MLLM localization but also improve the transparency and practical usability of deepfake forensic diagnosis.

F Implementation and Serialization Conventions

F.1 Stage Boundaries and Validity Masks

Feature extraction, evidence prediction, localization, and diagnosis have distinct trainable components. MARLIN and the Whisper encoder are frozen feature extractors. The shift adapter is fitted before Boundary-AEP and frozen thereafter. Boundary-AEP is optimized using the evidence objectives; the MLLM backbone remains frozen during the two separate localization LoRA adaptations. Diagnosis operates on already frozen localization results. BLA-AMR and Safe-AMR are not inference-time modules.

Dense feature timestamps and anchor timestamps are kept explicitly. Pooling and temporal resampling use these timestamps; padding does not create new observed intervals. The implementation-level feature widths are 768 for MARLIN ViT-B and 1,280 for Whisper large-v3 before temporal projection. These backbone widths are distinct from the shared $d_h = 512$ evidence representation. Missing features are represented with validity information and cannot be interpreted as low abnormality or evidence of authenticity.

Supervised language losses apply to assistant answer tokens only. System instructions, user text, media tokens, and padding are excluded from the target mask. Each localization branch is supervised only for its assigned modality. The visual counterfactual branch reuses the same visual target rather than changing a fake target to real when audio is withheld. Training schedules, sampling budgets, and optional controller switches belong to a frozen run configuration; an illustrative example or a library default does not establish the configuration of an experimental result.

F.2 Message Assembly and Output Serialization

The following pseudocode separates the model-independent prompt contract from the backbone's chat-template implementation. Media are supplied as multimodal inputs, not merely as filenames in the text. The audio and visual branches use the same clip; only the branch instruction, evidence view, and active LoRA adapter differ.

```
# Schematic assembly; media objects use the backbone processor API.
evidence = render_view(frozen_pack, modality)
text = fill_template(branch_prompt[modality], duration, evidence)
messages = [
    {"role": "system", "content": shared_system_instruction},
    {"role": "user", "content": [raw_audio_video, text]}
]
inputs = processor(messages, add_generation_prompt=True)
raw_answer = decode(inputs, adapter=localization_lora[modality])
parsed = validate_and_sanitize(raw_answer, modality, duration)
```

Input assembly uses an allowlist of media, duration, and inference-time evidence fields. Training targets, ground-truth intervals, dataset state labels, and controller statistics are not serialized into the localization prompt. Diagnostic prompts additionally receive the system-assembled frozen localization and provenance, never a mutable replacement for the localization output.

JSON serialization uses finite numeric values, arrays for interval lists, and `null` for explicitly unavailable optional values. Newlines inside a string must be escaped as `\n`; line wrapping in the typeset listings is only a display operation. Strings such as `RR | FR` in a prompt describe allowed choices: an actual answer must select one value, not emit the alternatives literally. An empty valid interval list, an invalid model response, and a withheld explanation remain distinct audit outcomes even when the evaluation protocol assigns no interval to a failed response.

All examples are labeled by provenance. A schema illustration explains the interface; a recorded input excerpt documents what was supplied to a branch; neither by itself establishes a decoded prediction or an experimental metric.