CAUTION: CALIBRATED UNCERTAINTY-GUIDED SAMPLE-ADAPTIVE EVIDENCE REASONING FOR SHORT-VIDEO FAKE NEWS DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

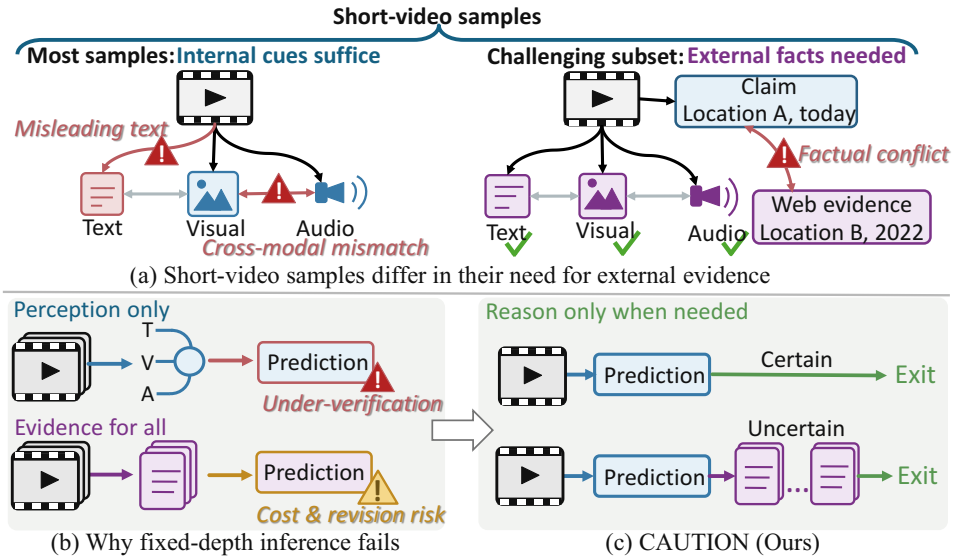
Short-video fake news samples vary substantially in verification difficulty: some can be reliably assessed from video-internal multimodal cues, whereas factually false yet internally coherent cases require external evidence. Existing methods nevertheless assign a fixed reasoning depth to all samples, either under-verifying evidence-dependent claims or unnecessarily subjecting otherwise reliable predictions to costly and potentially harmful slow deliberation. To address these challenges, we propose CAUTION, a calibrated uncertainty-guided framework for adaptive-depth evidence reasoning. First, a lightweight multimodal fast system exploits video-internal textual, visual, and acoustic cues to form an initial veracity estimate. Second, a calibrated adaptive-depth controller uses label-conditional conformal prediction to route uncertain samples to slow deliberation and trajectory-level calibration to determine when deliberation should stop. Third, a progressive relational evidence-revision module incrementally organizes ranked claim–evidence analysis units into a sample-specific heterogeneous graph, where typed support, refute, and unresolved relations capture their argumentative roles. Relation-aware graph attention aggregates the evolving evidence, while a reliability-gated residual revises rather than replaces the initial estimate. We further construct SV-CoRE, a large-scale claim-oriented evidence repository aligned with `FakeSV` and `FakeTT`, through retrieval, cleaning, deduplication, ranking, and large-model claim–evidence analysis. Experiments on both benchmarks demonstrate that CAUTION achieves state-of-the-art detection performance while substantially reducing unnecessary slow deliberation.

1 INTRODUCTION

Short-video fake news exhibits diverse forms of deception. Some false videos contain observable internal anomalies, such as manipulated content, misleading language, or inconsistencies among visual, textual, and acoustic modalities. Others are internally coherent but place authentic content in a false factual context, for example by misrepresenting an event's location, date, participants, or cause Bu et al. (2024); Qi et al. (2023a). The former can often be identified from video-internal cues, whereas the latter require external information to verify the claim conveyed by the video. Reliable detection must therefore coordinate multimodal perception with factual evidence rather than relying exclusively on either source.

Existing studies have mainly strengthened one of these capabilities. Compact multimodal detectors model cross-modal interactions and modality-specific inconsistencies, providing efficient predictions from the video itself Qi et al. (2023a); Bu et al. (2024). Large-language-model approaches introduce broader parametric knowledge and stronger generative reasoning Wang et al. (2025; 2026a). Recent work further incorporates externally retrieved knowledge into short-video analysis Wang et al. (2026b). Despite drawing on increasingly rich information sources, these methods generally assign every sample a predetermined inference path. They therefore do not determine, for each sample, whether external verification is needed or how many claim-wise evidence rounds are sufficient. Figure 1 illustrates this fixed-depth dilemma and the need for sample-adaptive reasoning.

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



(a) Short-video samples differ in their need for external evidence

(b) Why fixed-depth inference fails

(c) CAUTION (Ours)

Figure 1: Motivation for sample-adaptive evidence reasoning. Most videos can be assessed from internal cues, whereas a challenging subset requires external facts. Fixed-depth inference either under-verifies difficult samples or applies costly slow deliberation uniformly; CAUTION adapts the reasoning depth to each sample.

Moving from such fixed inference paths to sample-adaptive verification is nontrivial for three reasons. **First, video-internal cues are sufficient for some samples but not for others.** Some contain decisive cues, while others remain ambiguous without external evidence. The required reasoning depth therefore varies across samples. **Second, a short video may express multiple factual claims that contribute unequally to the final veracity judgment.** Verifying some claims may decisively confirm or overturn the video's central narrative, whereas resolving others may have little bearing on the final decision. Which claim is examined first can substantially affect verification efficiency. **Third, external evidence is fragmented and not uniformly beneficial.** Retrieved sources may be incomplete, redundant, or conflicting, and may support, refute, or remain inconclusive with respect to a claim. Treating such evidence as a fixed, unstructured context obscures its argumentative roles and the evolving state of verification, while irrelevant evidence or ungrounded analyses may even overturn an initially correct prediction.

Together, these challenges lead to the central question of this work: *How can a short-video detector determine, for each sample, whether external verification is needed, in what order its claims should be examined, and when the accumulated evidence is sufficient to retain or revise the initial prediction?*

To answer this question, we propose **CAUTION**, a calibrated uncertainty-guided framework for sample-adaptive evidence reasoning in short-video fake news detection. CAUTION couples a lightweight multimodal fast system with an evidence-grounded slow system. Every video is first represented as a Round-0 graph over its textual, visual, acoustic, and global-veracity representations, producing an initial veracity estimate. A label-conditional conformal controller accepts decisive fast predictions and routes only conformally uncertain samples to the slow system Shafer & Vovk (2008). During slow deliberation, trajectory-level calibration accounts for repeated checks as the controller is reapplied after each evidence round. Calibrated uncertainty therefore determines both whether slow deliberation is invoked and how deeply each sample should reason.

Once a video is routed to the slow system, CAUTION derives and prioritizes its checkable factual claims, then retrieves and ranks web evidence for the next scheduled claim at each round. An evidence reasoner evaluates the current claim against the retrieved sources, producing a claim-specific analysis and one of three claim–evidence relations: support, refute, or unresolved. This structured claim–evidence unit is incorporated into the sample-specific heterogeneous graph, progressively extending the Round-0 graph as $G^{(0)} \subset G^{(1)} \subset \cdots \subset G^{(r)}$. Relation-aware graph attention integrates the newly acquired evidence with the existing multimodal context, while a reliability-gated residual

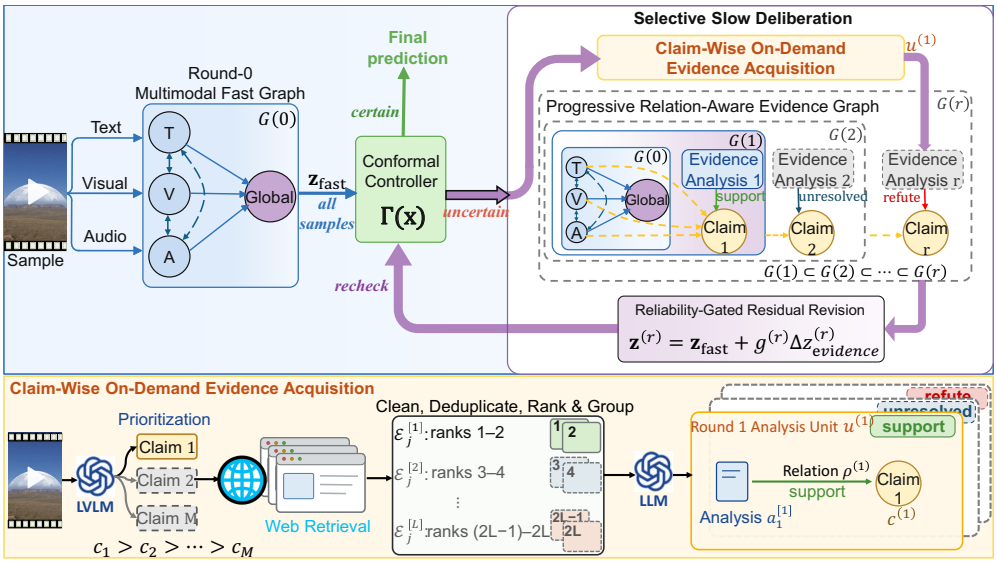


Figure 2: Overview of CAUTION. A conformal controller accepts decisive fast predictions and routes conformally uncertain videos to progressive claim-wise deliberation, rechecking the prediction after each graph update.

converts the graph update into a controlled revision of the fast logits. The revised logits are returned to the conformal controller, which either terminates verification under the calibrated stopping criterion or proceeds to the next scheduled evidence unit. This loop aligns claim selection, evidence acquisition, logit revision, and stopping for each sample.

We further construct **SV-CoRE** (**S**hort-**V**ideo **C**laim-**o**riented **R**epository of **E**vidence), containing 27,699 graph-ready units aligned with `FakeSV` and `FakeTT`. CAUTION achieves the highest Accuracy and Macro-F1 among the evaluated methods while routing fewer than half of the test samples to slow deliberation.

Our main contributions are threefold:

- **Sample-adaptive conformal control.** We propose CAUTION, which uses label-conditional conformal prediction to coordinate a lightweight multimodal fast system with evidence-grounded slow deliberation. The controller determines, for each sample, whether additional reasoning is required and when it should stop.
- **Prioritized claim-wise reasoning over progressive evidence graphs.** CAUTION extracts and prioritizes multiple factual claims, acquires evidence one claim at a time, and progressively expands a sample-specific graph with typed support, refute, and unresolved relations. Relation-aware attention and reliability-gated residual revision enable structured evidence accumulation without indiscriminately replacing the fast-system output.
- **Large-scale evidence repository and empirical validation.** We construct SV-CoRE with 27,699 graph-ready units in `FakeSV` and `FakeTT` partitions. Experiments show that CAUTION outperforms both fast-only and uniform all-evidence inference while routing fewer than half of the samples to slow deliberation.

2 The Proposed Model

2.1 Problem Formulation and Framework Overview

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i = (T_i, V_i, A_i)$ contains the textual, visual, and acoustic modalities of a short video and $y_i \in \mathcal{Y} = \{0, 1\}$ is its veracity label. Unlike fixed-depth detectors, CAUTION jointly produces a veracity prediction and determines a sample-specific deliberation depth:

$$(\hat{y}_i, r_i^*) = F_\theta(x_i), \qquad r_i^* \in \{0, \ldots, R_i\}, \tag{1}$$

where F_θ is the detector parameterized by θ, $\widehat{y}_i \in \mathcal{Y}$ is its final prediction, and r_i^* is the number of evidence rounds allocated to sample x_i under a maximum budget R_i. In particular, $r_i^* = 0$ denotes a fast-system exit, whereas $r_i^* > 0$ indicates selective slow deliberation.

As Figure 2 illustrates, the multimodal fast system predicts for every video. The label-conditional conformal controller accepts a prediction-consistent singleton and routes other cases to slow deliberation. For routed videos, prioritized claims schedule one evidence unit per round. Relation-aware graph attention updates the global veracity representation, and a reliability-gated residual revises the fast logits. After each round, the conformal singleton and calibrated trajectory-preservation conditions determine whether deliberation stops.

2.2 Multimodal Fast Graph and Conformal Routing

Round-0 multimodal fast graph. For video x, modality-specific adapters project textual, visual, and acoustic features into a shared space. Cross-modal self-attention yields $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{h}}_t; \widetilde{\mathbf{h}}_v; \widetilde{\mathbf{h}}_a]$, and a softmax modality gate assigns sample-dependent weights $\boldsymbol{\omega}$ over $\mathcal{M} = \{t, v, a\}$, forming $\mathbf{b} = \sum_{m\in\mathcal{M}} \omega_m \widetilde{\mathbf{h}}_m$. The Round-0 graph has a global veracity node, self-loops, and directed modality-to-global and cross-modal edges. Its encoder produces $\mathbf{h}_g^{(0)}$, which explicitly contributes to the fast logits:

$$\mathbf{z}_{\text{fast}} = \psi_b(\mathbf{b}) + \sum_{m\in\mathcal{M}} \omega_m \psi_m(\widetilde{\mathbf{h}}_m) + \psi_g(\mathbf{h}_g^{(0)} - \mathbf{b}). \tag{2}$$

The three terms combine fused, modality-level, and graph information; the global node thus affects the initial prediction even when no external evidence is used. Projection, attention, and gate details appear in Appendix A.

Conformal-guided selective routing. Uniform evidence deliberation is costly and may unnecessarily revise correct predictions. Therefore, CAUTION uses label-conditional conformal prediction to identify samples for which the fast prediction is not sufficiently decisive. For calibration sample (x_i, y_i),

$$\begin{aligned} \mathbf{p}^{(0)}(x) &= \mathrm{softmax}\big(\mathbf{z}_{\text{fast}}(x)/T_{\text{fast}}\big), \\ s_i^{(0)} &= 1 - p_{y_i}^{(0)}(x_i). \end{aligned} \tag{3}$$

For label ℓ, let $\mathcal{S}_\ell^{(0)} = \{s_i^{(0)} : y_i = \ell\}$ contain n_ℓ scores. Given an error level α, we compute

$$\begin{aligned} \mathcal{S}_\ell^{(0)+} &= \mathcal{S}_\ell^{(0)} \uplus \{1\}, \\ k_\ell &= \lceil (n_\ell + 1)(1-\alpha) \rceil, \\ \widehat{q}_\ell^{(0)} &= s_{\ell,(k_\ell)}^{(0)+}. \end{aligned} \tag{4}$$

The resulting prediction set is $\Gamma^{(0)}(x) = \{\ell : 1 - p_\ell^{(0)}(x) \le \widehat{q}_\ell^{(0)}\}$. Let $\widehat{y}_{\text{fast}} = \arg\max_\ell p_\ell^{(0)}(x)$. CAUTION accepts the fast prediction only if $\Gamma^{(0)}(x) = \{\widehat{y}_{\text{fast}}\}$; empty, multi-label, and prediction-inconsistent singleton sets are routed to slow deliberation. Under class-conditional exchangeability,

$$\Pr\big(Y \in \Gamma^{(0)}(X) \mid Y = \ell\big) \ge 1 - \alpha. \tag{5}$$

Since an incorrect fast exit implies that the true label is absent from $\Gamma^{(0)}(X)$, the class-conditional probability of an erroneous fast exit is also bounded by α. Thus, conformal uncertainty controls the reasoning path rather than serving only as a post-hoc confidence score.

2.3 Selective Deliberation over Progressive Evidence Graphs

Samples routed to slow deliberation enter an iterative evidence-grounded process. Starting from $G^{(0)}$, CAUTION acquires claim-level evidence on demand, progressively expands the graph, revises the fast logits, and reapplies the conformal criterion after each round.

Claim-wise on-demand evidence acquisition. An LVLM receives the complete short video together with its title and produces an ordered sequence of checkable factual claims $\mathcal{C}(x) = (c_1, \ldots, c_M)$ before any evidence is acquired. Claim-conditioned retrieval returns ranked candidates $e_{j,k}$ for claim c_j, which are filtered, deduplicated, and ranked by retrieval relevance and source quality. Adjacent ranks form evidence groups: CAUTION examines ranks 1–2 across claims before moving to deeper pairs (Appendix B). At round r, the large evidence reasoner returns $(\mathbf{a}^{(r)}, \rho^{(r)}) = F_{\mathrm{LER}}(c^{(r)}, \mathcal{E}^{(r)})$, where $\rho^{(r)} \in \{\mathrm{sup}, \mathrm{ref}, \mathrm{unr}\}$. The large model supplies structured evidence analysis rather than an independent final prediction.

Progressive relation-aware graph reasoning. The round-r unit $u^{(r)} = (c^{(r)}, \mathbf{a}^{(r)}, \rho^{(r)})$ adds or reuses a claim node, inserts an evidence-analysis node, and expands $G^{(r)} = G^{(r-1)} \oplus u^{(r)}$. Its edges distinguish structural roles and support, refute, or unresolved relations. On each edge, relation-specific embeddings modify the attention key and value, while a relation bias adjusts the attention weight. Aggregating the weighted values through residual and feed-forward layers yields the round-r global state $\mathbf{h}_g^{(r)}$; Appendix B gives the full update.

Reliability-gated residual revision. CAUTION isolates the evidence-induced graph change by comparing $\mathbf{h}_g^{(r)}$ with a matched counterfactual state $\mathbf{h}_{g,\mathrm{cf}}$: the latter masks all claim and evidence-analysis nodes and their incident edges but retains the multimodal subgraph and the same encoder. Let $\boldsymbol{\xi}^{(r)}$ collect the Round-0 and current global states, $\mathbf{h}_g^{(r)} - \mathbf{h}_{g,\mathrm{cf}}$, and the original modality weights. The learned gate is $g^{(r)} = \sigma(f_{\mathrm{gate}}(\boldsymbol{\xi}^{(r)}))$; its full input definition appears in Appendix B. Let $\boldsymbol{\delta}_g^{(r)}$ denote the transformed difference between $\mathbf{h}_g^{(r)}$ and $\mathbf{h}_{g,\mathrm{cf}}$, and $\boldsymbol{\delta}_b^{(r)}$ the transformed change in the modality representation after evidence-conditioned reweighting (Appendix B). The evidence correction and revised logits are

$$\begin{aligned} \Delta\mathbf{z}_{\mathrm{evi}}^{(r)} &= \boldsymbol{\delta}_g^{(r)} + \boldsymbol{\delta}_b^{(r)} + f_{\Delta}(\boldsymbol{\xi}^{(r)}), \\ \mathbf{z}^{(r)} &= \mathbf{z}_{\mathrm{fast}} + g^{(r)} \Delta\mathbf{z}_{\mathrm{evi}}^{(r)}. \end{aligned} \tag{6}$$

The gate thus scales the evidence correction before it is added to the fast logits, rather than replacing the initial prediction.

Trajectory-calibrated conformal stopping. To account for repeated inspection along the deliberation trajectory, each calibration sample is assigned its maximum slow-system nonconformity:

$$\begin{aligned} \mathbf{p}^{(r)}(x) &= \mathrm{softmax}\big(\mathbf{z}^{(r)}(x)/T_{\mathrm{slow}}\big), \\ s_i^{\mathrm{slow}} &= \max_{1 \le r \le R_i} \big[1 - p_{y_i}^{(r)}(x_i)\big]. \end{aligned} \tag{7}$$

Label-conditional thresholds $\widehat{q}_\ell^{\mathrm{slow}}$ use the same augmented order-statistic rule as fast routing, yielding $\Gamma^{(r)}(x) = \{\ell : 1 - p_\ell^{(r)}(x) \le \widehat{q}_\ell^{\mathrm{slow}}\}$. Under class-conditional exchangeability, trajectory calibration provides simultaneous coverage over the monitored rounds:

$$\Pr\Big(\exists r \le R(X) : Y \notin \Gamma^{(r)}(X) \mid Y = \ell\Big) \le \alpha. \tag{8}$$

Deliberation stops when either the prediction-consistent conformal singleton or the calibrated trajectory-preservation criterion accepts the current prediction; otherwise, the next scheduled unit is incorporated. Before returning the result, CAUTION retains a changed prediction only when it agrees with the accumulated relation directions. Full controller details are provided in the supplementary material.

2.4 Two-Stage Training

Stage I trains the multimodal fast predictor with $\mathcal{L}_{\mathrm{fast}} = N^{-1} \sum_i \mathrm{CE}(\mathbf{z}_{\mathrm{fast},i}, y_i)$, then freezes it as the Stage-II teacher. Stage II trains evidence revision on cumulative graph prefixes $\mathcal{P}_i \subseteq \{1, \ldots, R_i\}$, exposing each sample to different evidence depths. Its classification loss $\mathcal{L}_{\mathrm{evi}}$ gives greater weight to samples misclassified by the fast teacher.

Let $\mathcal{B}_+$ and $\mathcal{B}_-$ be the trajectory instances whose samples are correctly and incorrectly classified by the frozen fast teacher, respectively. On $\mathcal{B}_+$, $\mathcal{L}_{\text{pres}}$ averages $D_{\text{KL}}(\mathbf{p}_i^{\text{F}} \| \mathbf{p}_i^{(r)})$ to discourage drift from reliable fast predictions. On $\mathcal{B}_-$, $\mathcal{L}_{\text{cor}}$ averages $[\mu - z_{i,y_i}^{(r)} + z_{i,\widehat{y}_i^{\text{F}}}^{(r)}]_+$ to favor the true label over the incorrect fast prediction by margin μ. The gate loss $\mathcal{L}_{\text{gate}}$ supervises whether the ungated evidence correction reduces cross-entropy relative to the frozen teacher. The Stage-II objective is

$$\mathcal{L}_{\text{slow}} = \lambda_{\text{evi}}\mathcal{L}_{\text{evi}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \lambda_{\text{cor}}\mathcal{L}_{\text{cor}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}. \tag{9}$$

After optimization, we fit the fast and slow temperatures, calibrate the label-conditional thresholds, and then fit the trajectory-preservation controller with the prediction components fixed. Appendix D gives the full loss definitions and fitting sequence.

3 SV-CoRE: Evidence Repository Construction

To provide a reproducible evidence repository for progressive deliberation, we construct **SV-CoRE** for `FakeSV` and `FakeTT`. For each video, an LVLM extracts all checkable factual claims in verification-priority order. Claim-specific retrieval then materializes ranked results, using Bocha Search for Chinese `FakeSV` claims and Tavily Search for English `FakeTT` claims. Raw results are normalized through URL canonicalization, invalid-content filtering, and duplicate removal, then ordered using retrieval rank and available source-quality signals. The ranked results are partitioned by retrieval depth into small claim–evidence groups. Given only the claim and retrieved evidence, the large evidence reasoner converts each group into a structured analysis and a support, refute, or unresolved relation. Automated quality control removes malformed or duplicate units. Retrieval and evidence analysis use neither benchmark labels nor outcome-dependent metadata. Table 1 summarizes the coverage and 27,699 graph-ready units of SV-CoRE. These ranked units are available to the slow system; construction and quality-control details appear in the supplementary material.

Dataset	Coverage	Claims	Units	Support	Refute	Unresolved
`FakeSV`	99.4%	14,096	19,129	7,395	5,463	6,271
`FakeTT`	95.7%	5,314	8,570	2,604	1,690	4,276

Table 1: Core statistics of SV-CoRE. Coverage is the percentage of videos with retrieved evidence; Support, Refute, and Unresolved count graph-ready units.

4 Experiments

4.1 Experimental Setup

Datasets and protocol. We evaluate on the public `FakeSV` Qi et al. (2023a) and `FakeTT` Bu et al. (2024) benchmarks. Following the established benchmark protocol, we use chronological train/validation/test splits containing $2{,}536/546/542$ and $1{,}393/299/299$ videos, respectively.

Baselines and implementation. For public release and reproducible evaluation, SV-CoRE materializes the outputs of claim extraction, web retrieval, and Qwen evidence analysis. During evaluation, CAUTION follows its on-demand inference procedure: the fast system decides whether to deliberate, and each subsequent evidence request is served by its cached output, one unit at a time until stopping. Caching replaces live service responses, not CAUTION's routing or stopping decisions. Table 2 compares conventional multimodal and LLM-based or evidence-enhanced methods; "–" denotes unavailable results. We use `qwen3.6-plus` as the large evidence reasoner for both datasets. Retrieved evidence and large-model analyses are cached to ensure reproducibility and avoid web and API drift. Each validation split is stratified into disjoint, equally sized model-selection and calibration subsets: the former selects checkpoints, while the latter is reserved for calibration. Temperature scaling is fitted on training logits. We set the conformal error level to $\alpha = 0.05$. Architecture and training settings are detailed in the supplementary material.

Metrics. The main comparison reports Accuracy, Macro-F1, Precision, and Recall. Selective-reasoning analyses also report Trigger, the percentage of test samples entering the slow system; Avg. Units, cached Qwen claim–evidence analyses revealed per test sample, counting fast exits as

Model	FakeSV				FakeTT			
	Accuracy	Macro-F1	Precision	Recall	Accuracy	Macro-F1	Precision	Recall
Conventional multimodal								
HCFC-Hou (2019)	0.7491	0.7361	0.7558	0.7330	0.7324	0.7200	0.7199	0.7465
HCFC-Media (2020)	0.7638	0.7583	0.7614	0.7566	0.6254	0.6223	0.6558	0.6684
TikTec (2021)	0.7343	0.7326	0.7323	0.7353	0.6622	0.6508	0.6584	0.6787
CAFE (2022)	0.7878	0.7830	0.7812	0.7860	0.6287	0.6246	0.6532	0.6689
FANVM (2021)	0.7952	0.7881	0.7981	0.7846	0.7157	0.7021	0.7022	0.7263
SVFEND (2023a)	0.8088	0.8054	0.8017	0.8062	0.7714	0.7563	0.7512	0.7756
MultiEMO (2023)	0.8258	0.8205	0.8230	0.8187	–	–	–	–
SVFEND-SNEED (2023b)	0.8266	0.8167	0.8465	0.8103	–	–	–	–
FakingRecipe (2024)	0.8229	0.8192	0.8306	0.8229	0.7793	0.7854	0.8165	0.7793
LLM-based and evidence-enhanced								
GPT-4 (2023)	0.6743	0.6734	0.7076	0.7567	0.6145	0.6066	0.6328	0.6531
Qwen-VL (2023)	0.7417	0.7635	0.7847	0.7434	0.5886	0.7389	0.6421	**0.8700**
CSI (2026b)	0.8321	0.8256	0.8399	0.8207	–	–	–	–
OpEvFake (2024)	0.8564	0.8494	0.8546	0.8452	0.8060	0.7930	0.7860	0.8142
D^2et (2026)	0.8579	0.8539	0.8608	**0.8706**	0.8395	0.8231	0.8170	0.8316
CAUTION (Ours)†	**0.8708**	**0.8670**	**0.8754**	0.8630	**0.8428**	**0.8264**	**0.8206**	0.8340

Table 2: Overall performance on FakeSV and FakeTT. Best results are shown in bold. †Our results are averaged over three random seeds. CSI is a local adaptation using the same frozen FakeSV evidence pool; its best run is reported.

Dataset	Setting	ACC	Macro-F1	Trigger (%)	Avg. Units	Reduction (%)	Correct/Hurt	Net
FakeSV	Fast only	0.8026	0.7974	0.00	0.000	100.00	0/0	0
	Always all evidence	0.8487	0.8464	100.00	5.935	0.00	76/51	25
	Ours	**0.8708**	**0.8670**	48.71	2.581	56.51	**49/12**	**37**
FakeTT	Fast only	0.8361	0.8173	0.00	0.000	100.00	0/0	0
	Always all evidence	0.8328	0.8165	100.00	4.599	0.00	3/4	-1
	Ours	**0.8428**	**0.8264**	45.82	1.047	77.24	**3/1**	**2**

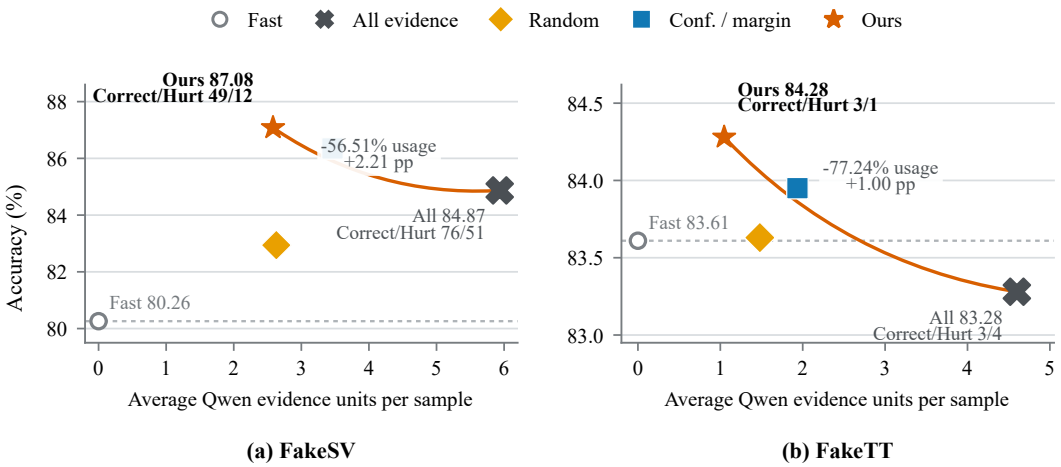
Table 3: Efficiency and Qwen usage. Reduction is usage saved versus Always all evidence.

zero; and Correct/Hurt, fast-system errors corrected and initially correct predictions harmed. Net is Correct minus Hurt; pp denotes percentage points.

4.2 Overall Performance

Conventional multimodal methods. CAUTION improves the best conventional Accuracy/Macro-F1 by 4.42/4.65 pp on FakeSV and 6.35/4.10 pp on FakeTT. Conventional detectors infer veracity primarily from video-internal cues, which capture visible manipulation and cross-modal inconsistency but cannot verify an internally coherent video placed in a false factual context. The gains suggest that CAUTION benefits from keeping multimodal perception as its default and adding claim-level verification when the fast prediction is uncertain.

LLM-based and evidence-enhanced methods. Direct GPT-4 and Qwen-VL predictions remain below task-specific models, showing that general-purpose multimodal reasoning alone is insufficient for fine-grained veracity discrimination. OpEvFake and D^2et instead use LLMs as semantic priors: the former extracts implicit credibility opinions for cross-modal diffusion, while the latter generates global-insight and local-veracity tags to guide multimodal denoising. CAUTION remains strongest on both benchmarks, indicating that factual grounding contributes information beyond such internal semantic enhancement. CSI is a baseline designed to retrieve online evidence, but applies the same multi-agent retrieval-and-review pipeline to every sample. CAUTION instead verifies only conformally uncertain samples and progressively incorporates typed, claim-level evidence, outperforming the local CSI adaptation on FakeSV. These results favor selective, structured evidence integration over a uniform retrieval-and-reasoning pipeline.



(a) FakeSV **(b) FakeTT**

Figure 3: Accuracy–usage trade-off. Average usage is measured in Qwen evidence units per test sample; marker area encodes trigger rate. Correct/Hurt counts corrections and harmful revisions.

Dataset	Variant	ACC	Macro-F1	Trigger (%)	Avg. Units	Correct/Hurt	Net
FakeSV	Full model	**0.8708**	**0.8670**	48.71	2.581	49/12	**37**
	w/o slow system	0.8026	0.7974	0.00	0.000	0/0	0
	w/o conformal routing	0.8487	0.8464	100.00	5.935	–	–
	w/o relation-aware edges	0.8635	0.8609	48.71	2.900	43/11	32
	w/o gated residual revision	0.8672	0.8632	48.71	3.063	50/15	35
	w/o preservation loss	0.8635	0.8601	48.71	2.351	46/13	33
FakeTT	Full model	**0.8428**	**0.8264**	45.82	1.047	3/1	**2**
	w/o slow system	0.8361	0.8173	0.00	0.000	0/0	0
	w/o conformal routing	0.8328	0.8165	100.00	4.599	–	–
	w/o relation-aware edges	0.8361	0.8181	45.82	1.813	1/0	1
	w/o gated residual revision	**0.8428**	**0.8264**	45.82	1.813	3/1	**2**
	w/o preservation loss	0.8261	0.8056	45.82	1.803	1/4	-3

Table 4: Component ablation on FakeSV and FakeTT. Correct/Hurt counts corrections and harmful revisions.

4.3 Selective Deliberation and Large-Model Usage

Figure 3 and Table 3 show that this procedure invokes 56.51% and 77.24% fewer Qwen analysis units than Always all evidence on FakeSV and FakeTT, respectively, while achieving higher Accuracy.

The correction counts clarify this trade-off. On FakeSV, Always all evidence corrects more fast-system errors, but its 76 corrections are offset by 51 harmful revisions, yielding Net $= 25$; CAUTION obtains a larger Net of 37 by limiting Hurt to 12. On FakeTT, uniform evidence use is harmful overall (Net $= -1$), whereas selective deliberation remains positive (Net $= 2$).

4.4 Component Ablation

Evidence allocation. In Table 4, *w/o slow system* lowers Accuracy by 6.82 pp on FakeSV and 0.67 pp on FakeTT. At the other extreme, *w/o conformal routing* sends every sample through all available evidence. Despite using more units, this variant trails the full model on both datasets and even falls below the fast-only baseline on FakeTT. Thus, the gain depends on allocating evidence, not merely adding it.

Evidence revision. *w/o relation-aware edges* collapses support, refute, and unresolved relations, weakening both datasets and showing that evidence direction matters after retrieval. In the *w/o gated residual revision* variant, Hurt rises from 12 to 15 on FakeSV; on FakeTT, the same Accuracy requires 1.813 rather than 1.047 average units. *w/o preservation loss* is most damaging on FakeTT,

Dataset	Trigger policy	ACC	Macro-F1	Trigger (%)	Avg. Units	Correct/Hurt	Net
FakeSV	Fast only	0.8026	0.7974	0.00	0.000	0/0	0
	Always all evidence	0.8487	0.8464	100.00	5.935	76/51	25
	Random trigger (mean)	0.8294	0.8255	52.58	2.630	29.6/15.1	14.6
	Max-softmax confidence trigger	0.8635	0.8603	54.43	3.461	52/19	33
	Ours: conformal trigger	**0.8708**	**0.8670**	48.71	2.581	49/12	**37**
FakeTT	Fast only	0.8361	0.8173	0.00	0.000	0/0	0
	Always all evidence	0.8328	0.8165	100.00	4.599	3/4	-1
	Random trigger (mean)	0.8363	0.8187	57.19	1.480	1.5/1.4	0.1
	Max-softmax confidence trigger	0.8395	0.8231	53.85	1.936	3/2	1
	Ours: conformal trigger	**0.8428**	**0.8264**	45.82	1.047	3/1	**2**

Table 5: Comparison of slow-deliberation trigger policies. Heuristics target a similar trigger rate; Random averages 20 draws. Max-softmax and probability-margin triggers are equivalent in binary classification. Correct/Hurt counts corrections and harmful revisions.

where Net changes from 2 to -3. Typed relations guide how evidence is interpreted, while the gate and preservation loss limit harmful changes to reliable fast predictions.

4.5 Why Conformal Routing?

Table 5 compares routing policies under similar budgets. Random routing is weaker, so routing frequency alone does not explain the gain. Max-softmax confidence routes more samples than conformal routing on both datasets yet yields lower Accuracy and Net. A single confidence threshold ranks samples by one score, whereas label-conditional conformal sets assess each label against its own calibrated threshold. This distinction helps explain why conformal routing achieves the highest Net on both benchmarks.

5 Related Work

Multimodal short-video fake news detection. FakeSV introduced SV-FEND to model cross-modal correlations and social context Qi et al. (2023a); subsequent methods characterize material selection and editing Bu et al. (2024), perform counterfactual inference Zong et al. (2025), or distill large-model reasoning into perception–cognition representations Wang et al. (2026a). These methods enrich video-internal representations. CSI adds external retrieval and multi-agent review Wang et al. (2026b), but applies a uniform deliberation pipeline to every input. CAUTION instead allocates claim-level evidence reasoning by calibrated uncertainty.

Evidence-grounded and adaptive reasoning. Automated fact-checking improves grounding through multi-source retrieval Li et al. (2025), graph-structured evidence modeling Chen et al. (2025), and iterative retrieval–verification Xie et al. (2025). FIRE is closest in adapting search depth, but targets textual claims and relies on model confidence. Selective classification abstains on uncertain inputs El-Yaniv & Wiener (2010), conformal prediction constructs distribution-free prediction sets Shafer & Vovk (2008), and test-time scaling allocates more computation to difficult instances Snell et al. (2025). CAUTION uses conformal uncertainty to control slow-system routing and depth, while typed claim–evidence relations progressively revise the multimodal graph.

6 Conclusion

Short-video fake news detection does not require a uniform reasoning depth. We present CAUTION, which allocates evidence reasoning according to each video's uncertainty. A multimodal fast system produces an initial estimate; label-conditional conformal routing decides whether to deliberate, and trajectory calibration with prediction preservation governs stopping. Prioritized claims progressively expand a sample-specific evidence graph, whose gated residual revises the fast prediction. SV-CoRE supplies 27,699 graph-ready units across FakeSV and FakeTT. CAUTION outperforms fast-only and all-evidence inference while routing fewer than half of test samples to slow deliberation and consuming fewer evidence-analysis units than uniform evidence use.

Reproducibility statement

Section 2 and Appendices A–D describe the model, calibration, stopping rule, and training objectives. Section 3 and Appendix E document SV-CoRE construction, while Section 4.1 and Appendix F specify the benchmark splits, implementation settings, and evaluation metrics. An anonymized archive containing the code and SV-CoRE is provided as supplementary material.

AI use statement

Generative AI tools were used to polish portions of the manuscript. We reviewed all AI-assisted work, checking the revised prose against the final methods and reported results. We take responsibility for the final content of this work.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. FakingRecipe: Detecting fake news on short video platforms from the perspective of creative process. In *ACM International Conference on Multimedia*, 2024.

Yingjian Chen, Haoran Liu, Yinhong Liu, Jinxiang Xie, Rui Yang, Han Yuan, Yanran Fu, Peng Yuan Zhou, Qingyu Chen, James Caverlee, and Irene Li. GraphCheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 14976–14995, 2025.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the Web Conference*, pp. 2897–2905, 2022.

Hyewon Choi and Youngjoong Ko. Using topic modeling and adversarial neural networks for fake news video detection. In *ACM International Conference on Information & Knowledge Management*, pp. 2950–2954, 2021.

Xueshan Deng, Penghang Yu, Shengsheng Qian, Baoyuan Qi, and Changsheng Xu. Denoising-enhanced multimodal detection network for fake news video detection. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. Towards automatic detection of misinformation in online medical videos. In *International Conference on Multimodal Interaction*, pp. 235–243, 2019.

Dayang Li, Fanxiao Li, Bingbing Song, Li Tang, and Wei Zhou. IMRRF: Integrating multi-source retrieval and redundancy filtering for LLM-based fake news detection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 9127–9142, 2025.

OpenAI, Josh Achiam, et al. GPT-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14444–14452, 2023a.

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. Two heads are better than one: Improving fake news video detection by correlating with neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11947–11959, 2023b.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *The 1st Workshop on NLP for COVID-19, Association for Computational Linguistics*, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A multimodal misinformation detector for COVID-19 short videos on TikTok. In *2021 IEEE International Conference on Big Data*, pp. 899–908, 2021.

Tao Shi and Shao-Lun Huang. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics*, pp. 14752–14766, 2023.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

JunXi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. FakeSV-VLM: Taming VLM for detecting fake short-video news via progressive mixture-of-experts adapter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 4782–4798, 2025.

Qingyan Wang, Lianwei Wu, Botao Wang, Wangkang, and Yaxiong Wang. From form to logic: Masked reconstruction and reasoning distillation for short video fake news detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 12698–12711, 2026a.

Yuxin Wang, Yang Yang, and Huaiwen Zhang. CSI: An investigative multi-agent framework for explainable short video fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25508–25528, 2026b.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. FIRE: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2901–2914, 2025.

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10817–10826, 2024.

Linlin Zong, Wenmin Lin, Jiahui Zhou, Xinyue Liu, Xianchao Zhang, Bo Xu, and Shimin Wu. Text-guided fine-grained counterfactual inference for short video fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1237–1245, 2025.

A Fast-System and Conformal-Routing Details

A.1 Round-0 Multimodal Graph

For a short-video sample x, modality-specific adapters project text, visual, and audio features into a shared space:

$$\mathbf{h}_m = \phi_m(\mathbf{f}_m) \in \mathbb{R}^d, \qquad m \in \mathcal{M} = \{t, v, a\}. \tag{10}$$

Cross-modal self-attention and layer normalization then produce

$$\begin{aligned} \mathbf{H} &= [\mathbf{h}_t; \mathbf{h}_v; \mathbf{h}_a], \\ \widetilde{\mathbf{H}} &= \mathrm{LN}(\mathbf{H} + \mathrm{MHA}(\mathbf{H}, \mathbf{H}, \mathbf{H})), \end{aligned} \tag{11}$$

where $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{h}}_t; \widetilde{\mathbf{h}}_v; \widetilde{\mathbf{h}}_a]$. A sample-dependent modality gate forms the fused multimodal representation:

$$\begin{aligned} \boldsymbol{\omega} &= \mathrm{softmax}\Big(g_{\mathrm{mod}}(\mathrm{vec}(\widetilde{\mathbf{H}}))\Big), \\ \mathbf{b} &= \sum_{m \in \mathcal{M}} \omega_m \widetilde{\mathbf{h}}_m. \end{aligned} \tag{12}$$

These representations initialize $G^{(0)} = (\mathcal{V}^{(0)}, \mathcal{A}^{(0)})$ with one global-veracity node and three modality nodes. Its typed topology contains self-loops, modality-to-global edges, and directed cross-modal edges. A lightweight typed graph-attention encoder returns the global state $\mathbf{h}_g^{(0)}$. The fast logits combine modality-level decisions with a graph-induced residual:

$$\begin{aligned} \mathbf{z}_{\mathrm{mm}} &= \psi_b(\mathbf{b}) + \sum_{m \in \mathcal{M}} \omega_m \psi_m(\widetilde{\mathbf{h}}_m), \\ \mathbf{z}_{\mathrm{fast}} &= \mathbf{z}_{\mathrm{mm}} + \psi_g(\mathbf{h}_g^{(0)} - \mathbf{b}). \end{aligned} \tag{13}$$

A.2 Label-Conditional Conformal Routing

Temperature scaling is fitted before conformal calibration and then held fixed. For $(x_i, y_i) \in \mathcal{D}_{\mathrm{cal}}$,

$$\begin{aligned} \mathbf{p}^{(0)}(x) &= \mathrm{softmax}(\mathbf{z}_{\mathrm{fast}}(x)/T_{\mathrm{fast}}), \\ s_i^{(0)} &= 1 - p_{y_i}^{(0)}(x_i). \end{aligned} \tag{14}$$

For label ℓ, let $\mathcal{S}_\ell^{(0)} = \{s_i^{(0)} : y_i = \ell\}$ contain n_ℓ scores. The augmented finite-sample quantile is

$$\begin{aligned} \mathcal{S}_\ell^{(0)+} &= \mathcal{S}_\ell^{(0)} \uplus \{1\}, \\ k_\ell &= \lceil (n_\ell + 1)(1 - \alpha) \rceil, \\ \widehat{q}_\ell^{(0)} &= s_{\ell,(k_\ell)}^{(0)+}, \end{aligned} \tag{15}$$

where $s_{\ell,(k_\ell)}^{(0)+}$ is the k_ℓ-th order statistic. The prediction set is

$$\Gamma^{(0)}(x) = \left\{ \ell \in \mathcal{Y} : 1 - p_\ell^{(0)}(x) \le \widehat{q}_\ell^{(0)} \right\}. \tag{16}$$

Writing $\widehat{y}_{\mathrm{fast}}(x) = \arg\max_\ell p_\ell^{(0)}(x)$, the routing decision is

$$\mathcal{R}(x) = \begin{cases} \text{fast exit}, & \Gamma^{(0)}(x) = \{\widehat{y}_{\mathrm{fast}}(x)\}, \\ \text{slow deliberation}, & \text{otherwise}. \end{cases} \tag{17}$$

Thus, empty, multi-label, and prediction-inconsistent singleton sets enter the slow system. Under class-conditional exchangeability,

$$\Pr\Big(Y \in \Gamma^{(0)}(X) \mid Y = \ell\Big) \ge 1 - \alpha. \tag{18}$$

Because an incorrect fast exit implies $Y \notin \Gamma^{(0)}(X)$, its class-conditional probability is also bounded by α.

B Progressive Evidence Deliberation

B.1 Claim-Wise Evidence Acquisition

For each routed sample, an LVLM first extracts a priority-ordered sequence of checkable claims $\mathcal{C}(x) = (c_1, \ldots, c_M)$. Claim-conditioned retrieval returns ranked candidates $e_{j,k}$ for claim c_j. Adjacent candidates form a depth group

$$\mathcal{E}_j^{[q]} = \{e_{j,2q-1}, e_{j,2q}\}, \qquad q = 1, \ldots, L, \tag{19}$$

and the depth-before-claim schedule is

$$\Pi(x) = \big((c_j, \mathcal{E}_j^{[q]})\big)_{\substack{q=1,\ldots,L \\ j=1,\ldots,M}}. \tag{20}$$

Consequently, ranks 1–2 are inspected across claims before any deeper rank is used. At round r, the large evidence reasoner returns

$$\begin{aligned} (\mathbf{a}^{(r)}, \rho^{(r)}) &= F_{\mathrm{LER}}(c^{(r)}, \mathcal{E}^{(r)}), \\ \rho^{(r)} &\in \{\mathrm{sup}, \mathrm{ref}, \mathrm{unr}\}. \end{aligned} \tag{21}$$

It therefore supplies a structured analysis and relation type, not an independent final prediction.

B.2 Relation-Aware Graph Propagation

The unit $u^{(r)} = (c^{(r)}, \mathbf{a}^{(r)}, \rho^{(r)})$ expands the graph as

$$G^{(r)} = G^{(r-1)} \oplus u^{(r)}, \qquad G^{(0)} \subset G^{(1)} \subset \cdots \subset G^{(r)}. \tag{22}$$

The expanded graph contains global, modality, claim, and evidence-analysis nodes. Node-type embeddings encode their semantic roles, and retrieval-stage embeddings encode evidence depth. Directed edges include modality-to-global, modality-to-claim, claim-to-global, and typed evidence-to-claim relations.

For an edge $j \to i$ of type ρ_{ji} at graph layer l and attention head h,

$$\begin{aligned} \mathbf{q}_i^{(l,h)} &= \mathbf{W}_Q^h \mathbf{h}_i^{(l)}, \\ \mathbf{k}_{ij}^{(l,h)} &= \mathbf{W}_K^h \mathbf{h}_j^{(l)} + \mathbf{r}_{\rho_{ji}}^{K,h}, \\ \mathbf{v}_{ij}^{(l,h)} &= \mathbf{W}_V^h \mathbf{h}_j^{(l)} + \mathbf{r}_{\rho_{ji}}^{V,h}. \end{aligned} \tag{23}$$

Relation information also enters the attention bias:

$$\begin{aligned} \beta_{ij}^{(l,h)} &= \mathrm{softmax}_{j \in \mathcal{N}(i)} \left(\frac{(\mathbf{q}_i^{(l,h)})^\top \mathbf{k}_{ij}^{(l,h)}}{\sqrt{d_h}} + \gamma_{\rho_{ji}}^h \right), \\ \mathbf{m}_i^{(l,h)} &= \sum_{j \in \mathcal{N}(i)} \beta_{ij}^{(l,h)} \mathbf{v}_{ij}^{(l,h)}. \end{aligned} \tag{24}$$

Residual and feed-forward transformations produce the round-r global state $\mathbf{h}_g^{(r)}$.

B.3 Reliability-Gated Residual Revision

A matched counterfactual graph masks all claim and evidence-analysis nodes and their incident edges while retaining the multimodal subgraph and graph encoder. Let its global state be $\mathbf{h}_{g,\mathrm{cf}}$. The evidence-induced change and revision gate are

$$\begin{aligned} \Delta \mathbf{h}_g^{(r)} &= \mathbf{h}_g^{(r)} - \mathbf{h}_{g,\mathrm{cf}}, \\ \boldsymbol{\xi}^{(r)} &= [\mathbf{h}_g^{(0)}; \mathbf{h}_g^{(r)}; \Delta \mathbf{h}_g^{(r)}; \boldsymbol{\omega}], \\ g^{(r)} &= \sigma\Big(f_{\mathrm{gate}}(\boldsymbol{\xi}^{(r)}) \Big). \end{aligned} \tag{25}$$

Evidence can also reweight the original modalities:

$$\begin{aligned} \boldsymbol{\omega}^{(r)} &= \mathrm{softmax}\Big(\log(\boldsymbol{\omega} + \epsilon) + g^{(r)} f_{\mathrm{route}}(\boldsymbol{\xi}^{(r)}) \Big), \\ \widetilde{\mathbf{b}}^{(r)} &= \sum_{m \in \mathcal{M}} \omega_m^{(r)} \widetilde{\mathbf{h}}_m. \end{aligned} \tag{26}$$

The graph and modality corrections are

$$\begin{aligned} \boldsymbol{\delta}_g^{(r)} &= \psi_{\mathrm{rev}}(\mathbf{h}_g^{(r)}) - \psi_{\mathrm{rev}}(\mathbf{h}_{g,\mathrm{cf}}), \\ \boldsymbol{\delta}_b^{(r)} &= \psi_{\mathrm{rev}}(\widetilde{\mathbf{b}}^{(r)}) - \psi_{\mathrm{rev}}(\mathbf{b}), \end{aligned} \tag{27}$$

and the residual update is

$$\begin{aligned} \Delta \mathbf{z}_{\mathrm{evi}}^{(r)} &= \boldsymbol{\delta}_g^{(r)} + \boldsymbol{\delta}_b^{(r)} + f_\Delta(\boldsymbol{\xi}^{(r)}), \\ \mathbf{z}^{(r)} &= \mathbf{z}_{\mathrm{fast}} + g^{(r)} \Delta \mathbf{z}_{\mathrm{evi}}^{(r)}. \end{aligned} \tag{28}$$

The slow system therefore controls how strongly accumulated evidence revises the original fast logits.

B.4 Trajectory Calibration and Stopping

After each round,

$$\begin{aligned} \mathbf{p}^{(r)}(x) &= \mathrm{softmax}\Big(\mathbf{z}^{(r)}(x)/T_{\mathrm{slow}}\Big), \\ s_i^{\mathrm{slow}} &= \max_{1\le r\le R_i}\Big[1-p_{y_i}^{(r)}(x_i)\Big]. \end{aligned} \tag{29}$$

For $\mathcal{S}_\ell^{\mathrm{slow}} = \{s_i^{\mathrm{slow}} : y_i = \ell\}$ of size n_ℓ^{slow}, we use

$$\begin{aligned} k_\ell^{\mathrm{slow}} &= \big\lceil (n_\ell^{\mathrm{slow}}+1)(1-\alpha)\big\rceil, \\ \widehat{q}_\ell^{\mathrm{slow}} &= \big[\mathcal{S}_\ell^{\mathrm{slow}} \uplus \{1\}\big]_{(k_\ell^{\mathrm{slow}})}. \end{aligned} \tag{30}$$

The round-r prediction set is

$$\Gamma^{(r)}(x) = \Big\{\ell : 1-p_\ell^{(r)}(x) \le \widehat{q}_\ell^{\mathrm{slow}}\Big\}. \tag{31}$$

Under class-conditional exchangeability, trajectory calibration gives

$$\Pr\Big(\exists r \le R(X) : Y \notin \Gamma^{(r)}(X) \mid Y=\ell\Big) \le \alpha. \tag{32}$$

Writing $\widehat{y}^{(r)} = \arg\max_\ell p_\ell^{(r)}(x)$, the conformal reference depth is

$$r^*_{\mathrm{conf}} = \min\Big(\Big\{r : \Gamma^{(r)}(x) = \{\widehat{y}^{(r)}\}\Big\} \cup \{R(x)\}\Big). \tag{33}$$

At inference, this criterion and the calibrated trajectory-preservation controller below are evaluated at each eligible prefix. If neither accepts the current prediction, the next unit in $\Pi(x)$ is incorporated. Before returning a changed label, CAUTION also checks its agreement with the accumulated relation directions: a change toward real requires at least as many support as refute relations, and a change toward fake requires at least as many refute as support relations. Otherwise, the fast prediction is retained.

C Calibrated Trajectory-Preservation Controller

CAUTION combines the trajectory-calibrated conformal singleton condition with a calibrated trajectory-preservation controller. The controller is fitted after the fast and slow prediction models and their conformal thresholds are fixed. For each routed training sample and evidence prefix r, we construct a 43-dimensional state vector $\boldsymbol{\phi}^{(r)}$ from predictive margins and entropies, changes along the trajectory, revision-gate and graph-state statistics, relation composition, modality weights, and relative evidence depth. Let $\widetilde{y}_i^{(r)}$ be the relation-consistent prediction at prefix r, and let $\widetilde{y}_i^{\mathrm{ref}}$ be the decision returned by the trajectory-calibrated conformal policy. A lightweight classifier h_θ is trained with target

$$t_i^{(r)} = \mathbb{I}\Big[\widetilde{y}_i^{(r)} = \widetilde{y}_i^{\mathrm{ref}}\Big]. \tag{34}$$

To calibrate the preservation condition, each held-out trajectory receives the maximum controller score among its unsafe prefixes,

$$u_i = \max_{r:\, t_i^{(r)}=0} h_\theta(\boldsymbol{\phi}_i^{(r)}), \tag{35}$$

with $u_i = 0$ when no unsafe prefix exists. Applying the finite-sample order-statistic rule to $\{u_i\}$ at risk level α_{stop} yields $\widehat{q}_{\mathrm{stop}}$, and the operating threshold is

$$\tau_{\mathrm{stop}} = \max\left(\widehat{q}_{\mathrm{stop}}, \tau_{\min}\right). \tag{36}$$

At inference, the conformal singleton condition and the preservation condition $h_\theta(\boldsymbol{\phi}^{(r)}) > \tau_{\mathrm{stop}}$ are evaluated after each eligible prefix. Deliberation stops when either condition accepts the current prediction; otherwise, the next evidence unit is incorporated. We use $\alpha_{\mathrm{stop}} = 0.01$ and the confidence floor $\tau_{\min} = 0.99$. Controller selection is performed on the model-selection split with all prediction components frozen.

D Training Objectives

D.1 Fast-System Learning

Given $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{N}$, Stage I jointly optimizes the modality adapters, cross-modal attention, modality gate, Round-0 graph encoder, and fast prediction heads:

$$\mathcal{L}_{\mathrm{fast}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}\left(\mathbf{z}_{\mathrm{fast},i}, y_i\right). \tag{37}$$

The resulting fast predictor is then frozen as the teacher for evidence-revision learning.

D.2 Evidence-Revision Learning

Let $\mathbf{z}_i^{\mathrm{F}}$, $\mathbf{p}_i^{\mathrm{F}}$, and $\widehat{y}_i^{\mathrm{F}}$ denote the teacher logits, probabilities, and prediction. For each sample, we use the complete trajectory and sampled intermediate prefixes $\mathcal{P}_i \subseteq \{1, \ldots, R_i\}$:

$$\mathcal{B}_{\mathrm{traj}} = \{(i, r) : r \in \mathcal{P}_i\}. \tag{38}$$

Initially misclassified samples receive a larger classification weight:

$$\begin{aligned} w_i &= 1 + (\lambda_{\mathrm{hard}} - 1)\mathbb{I}[\widehat{y}_i^{\mathrm{F}} \neq y_i], \\ \mathcal{L}_{\mathrm{evi}} &= \frac{1}{|\mathcal{B}_{\mathrm{traj}}|} \sum_{(i,r)\in\mathcal{B}_{\mathrm{traj}}} w_i \ \mathrm{CE}(\mathbf{z}_i^{(r)}, y_i). \end{aligned} \tag{39}$$

We partition trajectory instances according to whether the fast teacher is correct:

$$\begin{aligned} \mathcal{B}_+ &= \{(i, r) \in \mathcal{B}_{\mathrm{traj}} : \widehat{y}_i^{\mathrm{F}} = y_i\}, \\ \mathcal{B}_- &= \{(i, r) \in \mathcal{B}_{\mathrm{traj}} : \widehat{y}_i^{\mathrm{F}} \neq y_i\}. \end{aligned} \tag{40}$$

The preservation and correction objectives are

$$\begin{aligned} \mathcal{L}_{\mathrm{pres}} &= \frac{1}{|\mathcal{B}_+|} \sum_{(i,r)\in\mathcal{B}_+} D_{\mathrm{KL}}\left(\mathbf{p}_i^{\mathrm{F}} \,\|\, \mathbf{p}_i^{(r)}\right), \\ \mathcal{L}_{\mathrm{cor}} &= \frac{1}{|\mathcal{B}_-|} \sum_{(i,r)\in\mathcal{B}_-} \left[\mu - z_{i,y_i}^{(r)} + z_{i,\widehat{y}_i^{\mathrm{F}}}^{(r)}\right]_+, \end{aligned} \tag{41}$$

where $[u]_+ = \max(u, 0)$. The former discourages drift from the fast probability distribution on correctly classified samples, while the latter encourages evidence to overturn an incorrect prediction by margin μ.

Before gating, the full candidate correction is

$$\mathbf{z}_{\mathrm{cand},i}^{(r)} = \mathbf{z}_i^{\mathrm{F}} + \Delta\mathbf{z}_{\mathrm{evi},i}^{(r)}. \tag{42}$$

Its reliability target indicates whether this candidate reduces classification loss:

$$\begin{aligned} \tau_i^{(r)} &= \mathbb{I}\Big[\mathrm{CE}(\mathbf{z}_{\mathrm{cand},i}^{(r)}, y_i) < \mathrm{CE}(\mathbf{z}_i^{\mathrm{F}}, y_i)\Big], \\ \mathcal{L}_{\mathrm{gate}} &= \frac{1}{|\mathcal{B}_{\mathrm{traj}}|} \sum_{(i,r)\in\mathcal{B}_{\mathrm{traj}}} \mathrm{BCE}(g_i^{(r)}, \tau_i^{(r)}). \end{aligned} \tag{43}$$

The complete Stage-II objective is

$$\mathcal{L}_{\mathrm{slow}} = \lambda_{\mathrm{evi}}\mathcal{L}_{\mathrm{evi}} + \lambda_{\mathrm{pres}}\mathcal{L}_{\mathrm{pres}} + \lambda_{\mathrm{cor}}\mathcal{L}_{\mathrm{cor}} + \lambda_{\mathrm{gate}}\mathcal{L}_{\mathrm{gate}}. \tag{44}$$

After optimization, we fit the fast and slow temperatures and estimate their label-conditional conformal thresholds. The trajectory-preservation controller is fitted and calibrated last, with all prediction components fixed.

E SV-CoRE Details

Claim-oriented retrieval. For each video, an LVLM receives the complete video and title and extracts checkable factual claims in verification-priority order. SV-CoRE construction performs claim-specific retrieval for every claim, combining it with contextual anchors from the video. We use Bocha Search for Chinese `FakeSV` claims and Tavily Search for English `FakeTT` claims. Each record retains its query, rank, URL, title, textual content, source domain, and provider score when available.

Cleaning and depth-wise organization. Raw records undergo URL canonicalization, invalid-content filtering, and duplicate removal. Candidates are ordered by retrieval rank and, when available, provider confidence, source quality, and cleaning scores. We retain up to six results per `FakeSV` claim and eight per `FakeTT` claim. Adjacent ranks form the depth groups 1–2, 3–4, 5–6, and, for `FakeTT`, 7–8. This materializes a reproducible pool of ranked evidence units for progressive deliberation.

Analysis-enriched units. Given only a claim and its retrieved evidence, the large evidence reasoner produces analysis text and a verdict from {*support*, *refute*, *abstain*}, mapped to the support, refute, and unresolved graph relations. For traceability, the SV-CoRE also retains evidence-matching assessments, support and refute scores, evidence-sufficiency scores, and detected manipulation or evidence-gap types. Automated quality control removes malformed or duplicate units.

Claim-conditioned retrieval and evidence availability. Retrieval is central to CAUTION's verification: queries use video-derived claims and context to find information relevant to each claim. Neither the queries nor Qwen's claim–evidence analyses receive benchmark veracity labels or outcome-dependent metadata.

SV-CoRE covers 99.4% of `FakeSV` and 95.7% of `FakeTT` and contains $19{,}129$ and $8{,}570$ graph-ready units, respectively. During evaluation, the units are replayed in the sample-specific order $\Pi(x)$. Fast exits use none; routed samples reveal units one at a time until stopping.

Repository construction and on-demand use. SV-CoRE is constructed across the benchmark splits independently of CAUTION's routing decisions. It materializes claim extraction, web retrieval, and Qwen evidence analysis for public reuse and reproducible comparison. At inference, CAUTION first applies the fast system: fast exits use no evidence, while routed samples follow $\Pi(x)$ and access one claim–evidence analysis unit per round until stopping. During evaluation, SV-CoRE supplies the precomputed outputs of these steps instead of live services; the routing and stopping procedure remains on demand.

F Experimental Protocol and Metrics

Datasets and protocol. We evaluate on the public `FakeSV` and `FakeTT` benchmarks. Their official train/validation/test splits contain $2{,}536/546/542$ and $1{,}393/299/299$ videos, respectively. SV-CoRE supplies CAUTION's slow system. Because CSI's intermediate agent outputs are unavailable, our local adaptation maps the first three cached `FakeSV` evidence units to its released adjudicator branches and reports its best observed run. Other baselines retain their original inputs. Retrieved evidence and large-model analyses are cached to avoid web and API drift.

Baselines. We compare against conventional multimodal detectors (HCFC-Hou, HCFC-Media, TikTec, CAFE, FANVM, SVFEND, MultiEMO, SVFEND-SNEED, and FakingRecipe) and LLM-based or evidence-enhanced methods (GPT-4, Qwen-VL, OpEvFake, D^{2}et, and CSI). Unavailable results are denoted by "–".

Implementation details. The graph hidden size is 256, with four attention heads and two graph layers. Training uses AdamW, weight decay 10^{-4}, batch size 24; evaluation uses batch size 64. For `FakeSV`/`FakeTT`, the base and dynamic learning rates are $(10^{-4}, 10^{-4})$ and $(7.5 \times 10^{-5}, 1.5 \times 10^{-4})$, with dropout 0.20 and 0.15, respectively. We use `qwen3.6-plus` as the large evidence reasoner for both datasets. We stratify the original validation set into equally sized model-selection

Inference policy	ACC	Macro-F1	Avg. Units
Fixed 1 round	0.8266	0.8248	0.991
Fixed 2 rounds	0.8432	0.8401	1.958
Fixed 3 rounds	0.8339	0.8303	2.856
Fixed all evidence	0.8487	0.8464	5.935
CAUTION (adaptive)	**0.8708**	**0.8670**	2.581

Table 6: Uniform fixed-depth inference versus CAUTION on `FakeSV`. All policies use the same graph predictor and relation-consistency check. Avg. Units averages evidence-analysis units over all test samples. Fixed all evidence thus differs from the raw-prediction Always all evidence baseline in Table 3.

and calibration subsets: the former selects checkpoints, whereas the latter is reserved for calibration. Temperature scaling is fitted on training logits. We use $\alpha = 0.05$ for conformal prediction and $\alpha_{\text{stop}} = 0.01$ for the trajectory-preservation controller. Unless stated otherwise, we use typed evidence relations, C3D visual features, and the relation-consistency check.

Metrics. The main comparison reports Accuracy, Macro-F1, Precision, and Recall. Selective-reasoning analyses additionally report Trigger, the percentage of test samples entering the slow system; Avg. Units, the mean number of Qwen claim–evidence analysis units revealed per test sample, counting fast exits as zero; and Correct/Hurt, counts of corrections and harmful revisions. Net is Correct minus Hurt.

F.1 Uniform Fixed-Depth Comparison

To examine sample-adaptive evidence depth, we compare CAUTION with uniform policies that expose every video to a fixed number of evidence rounds. All variants use the same trained graph model and relation-consistency check; they differ in evidence allocation. If fewer units are available than the prescribed depth, all available units are used.

CAUTION achieves the highest Accuracy and Macro-F1 among these policies. On `FakeSV`, neither uniformly shallow nor exhaustive reasoning provides the same accuracy–usage trade-off as sample-adaptive evidence allocation.