LEARNING WHAT TO TRANSFER: PERMISSION-ROUTED MECHANISM EXPERTS FOR MULTIMODAL SHORT-VIDEO MISINFORMATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting misinformation in short videos requires reasoning over both internal multimodal content and external event evidence. Existing methods often fuse heterogeneous internal cues into a shared representation or directly incorporate retrieved precedents, overlooking differences among manipulation mechanisms and the transferability of external information. To address this issue, we propose Permission-Routed Mechanism Experts (PRME). PRME assigns soft Strong, Weak, and No permissions to retrieved precedents: Strong permits factual feature transfer and mechanism guidance, Weak retains only mechanism-level priors, and No relies solely on internal evidence. A shared pool of mechanism-specific experts then performs specialized authenticity reasoning. Experiments on FakeSV and FakeTT demonstrate that PRME consistently improves retrieval-augmented misinformation detection, achieving notable gains on the FakeSV temporal and event splits and the FakeTT temporal split. Ablation and relation-stratified analyses further validate the effectiveness of permission-aware evidence transfer and mechanism-specific expert routing.

1 INTRODUCTION

Online misinformation can distort public understanding, erode trust in information systems, and disrupt responses to emerging events. Short-video platforms have become important channels of information dissemination, where low production costs and rapid redistribution accelerate the spread of deceptive content. Unlike text- or image-based misinformation, short-video misinformation is often semantic and event-dependent rather than purely synthetic: authentic footage may be paired with a false context, such as an incorrect location, an outdated event, a replaced subject, or a distorted causal relation. Its detection therefore requires joint reasoning over multimodal content and the event claims conveyed by that content (Qi et al., 2023a; Bu et al., 2024).

Existing methods follow two complementary directions. **Internal-evidence methods** assess authenticity from the query video by integrating multimodal content and social context, with recent work modeling content heterogeneity, editing patterns, modality bias, rationales, and multiple detection views (Choi & Ko, 2021; Qi et al., 2023a; Bu et al., 2024; Zong et al., 2025; Hong et al., 2025; Li et al., 2026a). **External-evidence methods** retrieve event neighbors, debunking videos, historical precedents, or related narratives for cross-instance verification (Qi et al., 2023b; Chen et al., 2026; Li et al., 2026b). Despite their complementary strengths, current research still faces **uncontrolled evidence transfer** and **incomplete mechanism modeling**.

Uncontrolled evidence transfer. A detector may use a retrieved precedent without determining whether it is transferable or which information can be safely transferred. In Fig. 1(a), a query claiming that "twin sisters married two brothers" retrieves a video about a bride-price dispute. Although semantically related, the videos involve different people and events, so direct factual transfer may create invalid analogies. The precedent may nevertheless reveal abstract manipulation mechanisms, such as fabricated facts or subject mismatch. Existing methods rank, filter, or weight evidence using semantic similarity, retrieval confidence, quality, or sufficiency (Atanasova et al., 2022; Pan et al., 2025; Wei et al., 2025; Ding et al., 2026), but do not separate factual transfer from

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.

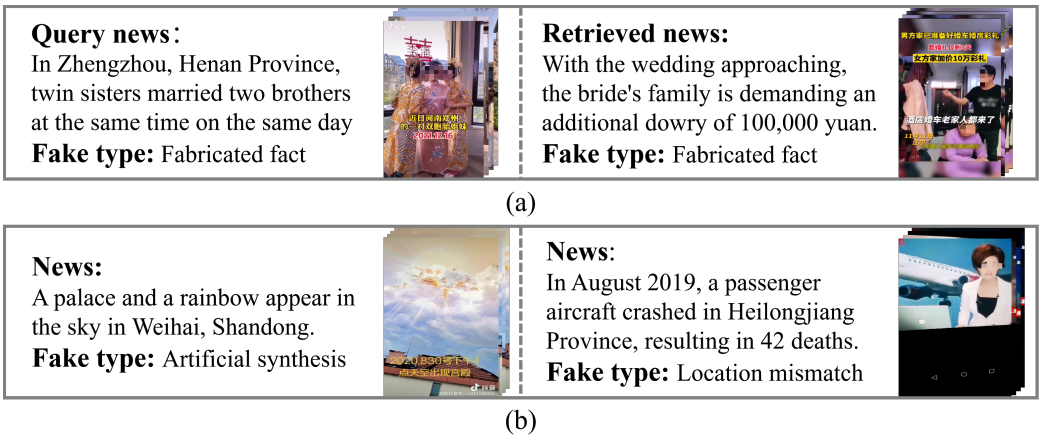


Figure 1: Illustrations of the two challenges. (a) The query and precedent share a broad marriage-related topic but involve different entities and event, making factual transfer unsafe. (b) Different misinformation mechanisms require mechanism-specific evidence and reasoning.

mechanism transfer. As a result, weak precedents are either injected with inadequate constraints or discarded despite their potentially useful mechanism-level cues.

Incomplete mechanism modeling. Different misinformation types require different evidence and reasoning. In Fig. 1(b), the "palace in the sky" case involves visual synthesis, whereas the aircraft-crash report involves a location mismatch. The former requires detecting manipulated visual content; the latter requires verifying geographic context. A shared representation can obscure these decisive cues and limit the use of relevant external evidence. Although prior work has modeled creation patterns (Bu et al., 2024) and multi-view evidence from content, emotion, and publisher profiles (Li et al., 2026a), it does not explicitly distinguish misinformation mechanisms or reason about each case accordingly.

Motivation: relation-dependent retrieval. On the FakeSV temporal test set, reproduced TrueLens reaches 96.30% Accuracy when a same-event precedent can be retrieved, but drops to 87.30% when only similar-event precedents are available and 87.80% for same-topic precedents. Importantly, weaker relations are not entirely uninformative: 93.91% of similar-event cases and all same-topic cases still have at least one retrieved reference sharing a manipulation mechanism with the target. This gap motivates us to isolate how retrieved information is used. Event-consistent precedents may support factual transfer, whereas weaker precedents should be prevented from altering query facts and used only for mechanism-level guidance when appropriate.

Based on this finding, we propose **Permission-Routed Mechanism Experts (PRME)**, comprising a hierarchical permission router, three permission-specific branches, and a pool of mechanism-aware experts. The router assigns soft STRONG, WEAK, or NO permissions. STRONG allows both factual representations and mechanism priors; WEAK blocks direct precedent-feature injection while retaining mechanism-level guidance; and NO removes external information and relies solely on internal evidence. PRME also includes eight specialist experts and a general expert for conservative fallback. PRME thus turns retrieval augmentation into permission-aware transfer and generic multimodal fusion into mechanism-specific reasoning.

Our contributions are threefold:

- We formulate retrieval-augmented short-video misinformation detection as event-relation-conditioned information transfer and distinguish factual transferability from mechanism transferability.
- We develop a permission-aware architecture with STRONG, WEAK, and NO routes that grant different access rights to external information, preventing weak precedents from altering the query's factual representation.

- We introduce mechanism-aware expert routing that converts weak precedents into constrained manipulation priors while preserving an internal-evidence fallback. On FakeSV, PRME outperforms the reproduced TrueLens baseline by 0.75 and 2.54 Macro-F1 points on the temporal and event splits, respectively. On FakeTT, it improves temporal Macro-F1 by 2.04 points and remains effectively tied under event isolation.

2 Related Work

Internal-evidence misinformation detection. Multimodal misinformation detection models authenticity through cross-modal relations, including event-invariant representations (Wang et al., 2018), joint text–image modeling (Singhal et al., 2019), and cross-modal consistency learning (Zhou et al., 2020; Chen et al., 2022). Short-video methods further integrate temporal visual content, audio, transcripts, social context, and publisher information, with recent work exploring topic-agnostic representations (Choi & Ko, 2021), multimodal–social fusion (Qi et al., 2023a), creation patterns (Bu et al., 2024), modality-bias mitigation (Zong et al., 2025), rationale-based reasoning (Hong et al., 2025), and multi-view aggregation (Li et al., 2026a). These methods largely fuse heterogeneous cues into a shared authenticity representation rather than explicitly routing cases by manipulation mechanism.

Retrieval-augmented verification. External evidence is useful when event context is absent from the query. NEED retrieves neighboring and debunking videos (Qi et al., 2023b), TrueLens incorporates real and fake precedents (Chen et al., 2026), and related work uses retrieved narratives for cross-instance verification (Li et al., 2026b). Other studies model evidence sufficiency, quality, adaptive use, or retrieval confidence (Atanasova et al., 2022; Pan et al., 2025; Wei et al., 2025; Ding et al., 2026). They mainly decide whether evidence should be selected or trusted, whereas PRME distinguishes factual transfer from mechanism-level guidance so that weakly related precedents can contribute abstract manipulation cues without overwriting query facts.

Expert specialization. Mixture-of-Experts models route heterogeneous inputs to specialized subnetworks (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022). Misinformation experts have been organized by domains (Nan et al., 2021), multimodal views (Ying et al., 2023), or text–video authenticity combinations (Wang et al., 2025). PRME instead organizes experts around manipulation mechanisms and couples expert routing with explicit permissions over external-evidence transfer.

3 Permission-Routed Mechanism Experts

3.1 Problem Formulation and Overview

Let q denote a short-video post with textual, visual, and audio inputs $\{x_q^m\}_{m\in\mathcal{M}}$, where $\mathcal{M} = \{t, v, a\}$, and authenticity label $y \in \{0, 1\}$ (0: authentic, 1: fake). For each evaluation partition, a unified precedent bank $\mathcal{B}$ is constructed exclusively from the training set. Query self-matches are removed, and no test sample or test-derived representation is inserted into $\mathcal{B}$.

Given q, PRME retrieves a candidate set $\mathcal{R}_q \subset \mathcal{B}$ and computes soft routing weights for three permission branches: Strong, Weak, and No. Strong permits direct feature interaction with the highest-ranked precedent and mechanism guidance; Weak blocks factual feature transfer but retains a precedent-derived mechanism prior; and No removes all precedent-derived inputs. Each branch constructs an expert input and expert-gating logits, all branches are evaluated by a shared mechanism-expert pool, and their predictions are fused by the permission weights.

3.2 Multimodal Representation and Retrieval

For each available modality m, a modality-specific encoder E_m followed by a feed-forward projection F_m produces a sequence representation $H_q^m \in \mathbb{R}^{L_m\times d}$ in a shared d-dimensional space. We mean-

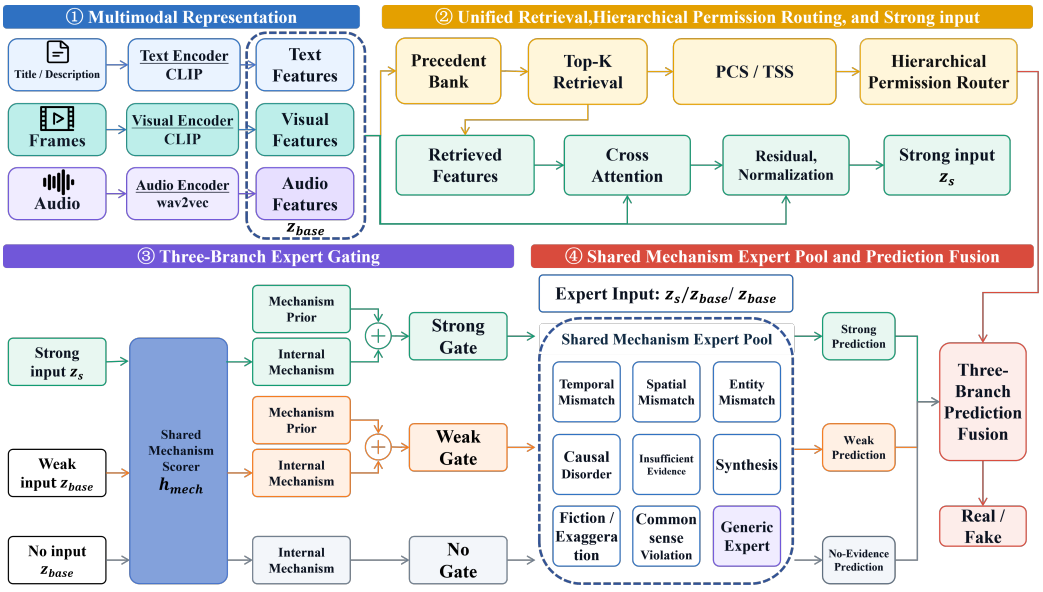


Figure 2: Overview of PRME. The router controls both branch contribution and information access: only Strong may inject retrieved precedent features, Weak receives mechanism priors without factual transfer, and No uses internal evidence only.

pool each sequence and average the available modalities:

$$H_q^m = F_m(E_m(x_q^m)), \tag{1}$$

$$\bar{h}_q^m = \mathrm{Pool}(H_q^m), \tag{2}$$

$$z_{\mathrm{base}} = \mathrm{LN}\left(\frac{1}{|\mathcal{M}_q|}\sum_{m\in\mathcal{M}_q}\bar{h}_q^m\right), \tag{3}$$

where $\mathcal{M}_q \subseteq \mathcal{M}$ is the set of available modalities, Pool is mean pooling, and LN is layer normalization. The same processing is applied to every precedent $r \in \mathcal{B}$, yielding H_r^m and $\bar{h}_r^m$.

For each modality, query–precedent similarity is $s_r^m = \cos(\bar{h}_q^m, \bar{h}_r^m)$. We rank precedents by a weighted multimodal retrieval score

$$s_r^{\mathrm{ret}} = \sum_{m\in\mathcal{M}_q}\eta_m s_r^m, \qquad \sum_{m\in\mathcal{M}_q}\eta_m = 1, \tag{4}$$

and take the top-k items as $\mathcal{R}_q$. Unlike designs that maintain separate authentic/fake memories, this unified training-only bank leaves the decision of *how* a retrieved item may be used to the permission router.

3.3 Hierarchical Permission Routing

A single retrieval score is insufficient for transfer control: surface/perceptual correspondence is more relevant to direct factual reuse, whereas semantic and event-level agreement can still support mechanism reasoning even when factual transfer is unsafe. For each $r \in \mathcal{R}_q$, PRME therefore computes perceptual consistency (PCS) and textual semantic support (TSS):

$$s_r^{\mathrm{pcs}} = \sum_{m\in\mathcal{M}_q}\gamma_m s_r^m, \tag{5}$$

$$s_r^{\mathrm{tss}} = \sum_{m\in\mathcal{M}_q}\beta_m s_r^m + \beta_e \delta_r^e, \tag{6}$$

where $\delta_r^e = 1$ when query and precedent share the dataset-provided $event_id$ and 0 otherwise. The modality coefficients and the $\beta_e = 0$ ablation are specified in Appendix C.3. We summarize the retrieved set by $P = \max_r s_r^{\mathrm{pcs}}$ and $S = \max_r s_r^{\mathrm{tss}}$.

Rather than a flat three-way softmax, PRME grants the strongest permission first and allocates only its residual mass to weaker routes:

$$p_s = \sigma((P - \tau_s)/T_r), \qquad \tilde{p}_w = \sigma((S - \tau_w)/T_r),$$
$$p_w = (1 - p_s)\tilde{p}_w, \qquad p_n = (1 - p_s)(1 - \tilde{p}_w). \tag{7}$$

Here τ_s and τ_w are the Strong/Weak thresholds, $T_r > 0$ is the routing temperature, and $\mathbf{p} = [p_s, p_w, p_n]$ sums to one. The hierarchy implements a conservative transfer order: direct factual transfer must pass the stricter perceptual gate; otherwise semantic support may grant mechanism-only transfer; the remaining mass falls back to internal evidence. The router is score-based and receives no direct authenticity-label supervision.

3.4 Permission-Specific Evidence Construction

Let $\mathcal{A} = \{1, \ldots, N\}$ index the manipulation mechanisms. Every training precedent stores a mechanism-affinity vector $\widetilde{\mathbf{a}}_r \in [0, 1]^N$ defined in Eq. equation 19. A shared mechanism scorer $h_{\text{mech}} : \mathbb{R}^d \to \mathbb{R}^N$, implemented as a three-layer MLP with ReLU activations and dropout, produces internal mechanism logits. The notation $[\cdot; \cdot]$ denotes concatenation, and b_0 is a learnable gate logit for the general expert. Each branch $c \in \{s, w, n\}$ outputs an expert input z_c and gate logits $\mathbf{u}_c \in \mathbb{R}^{N+1}$.

Strong: factual and mechanism transfer. Let r_1 be the highest-ranked precedent under Eq. equation 4. For each modality, the query attends to the corresponding precedent sequence:

$$\widetilde{H}_s^m = \mathrm{LN}\big(H_q^m + \alpha\, \mathrm{MHA}_m(H_q^m, H_{r_1}^m, H_{r_1}^m)\big)\,, \tag{8}$$

$$z_s = \mathrm{LN}\left(\frac{1}{|\mathcal{M}_q|} \sum_{m \in \mathcal{M}_q} \mathrm{Pool}(\widetilde{H}_s^m)\right), \tag{9}$$

$$\mathbf{u}_s = [h_{\text{mech}}(z_s) + \lambda_s \widetilde{\mathbf{a}}_{r_1};\, b_0]. \tag{10}$$

MHA_m is modality-specific multi-head attention and α controls precedent feature injection. Strong is the *only* branch in which precedent features can directly alter the query representation; λ_s controls the accompanying mechanism prior.

Weak: mechanism transfer without factual injection. Weak preserves the internal representation, $z_w = z_{\text{base}}$, and aggregates only precedent mechanism affinities using TSS-normalized weights:

$$\omega_r = \frac{\exp(s_r^{\text{tss}}/T_p)}{\sum_{r' \in \mathcal{R}_q} \exp(s_{r'}^{\text{tss}}/T_p)}, \tag{11}$$

$$\boldsymbol{\pi}_w = \mathrm{clip}\left(\sum_{r \in \mathcal{R}_q} \omega_r \widetilde{\mathbf{a}}_r, 0, 1\right), \tag{12}$$

$$\mathbf{u}_w = [h_{\text{mech}}(z_w) + \lambda_w \boldsymbol{\pi}_w;\, b_0]. \tag{13}$$

Here $T_p > 0$ is the prior-aggregation temperature and clipping is element-wise. Thus Weak precedents may indicate *how* a query could be manipulated without injecting their entities, locations, times, or causal sequence into the query features.

No: internal-evidence fallback. No discards all precedent-derived inputs:

$$z_n = z_{\text{base}}, \qquad \mathbf{u}_n = [\lambda_n h_{\text{mech}}(z_n);\, b_0 + \lambda_0]. \tag{14}$$

λ_n scales mechanism-expert logits, while λ_0 increases accessibility of the general expert when neither factual nor mechanism-level transfer is reliable.

3.5 Mechanism Experts and Prediction

The shared pool contains $N = 8$ mechanism experts for temporal mismatch, spatial mismatch, entity mismatch, causal disorder, insufficient evidence, synthesis, fabricated/exaggerated facts, and commonsense violation, plus a general expert E_0. Each $E_j : \mathbb{R}^d \to \mathbb{R}^2$ is an independently parameterized

feed-forward classifier head; h_{mech} is used to construct the gates rather than to directly classify authenticity. For branch $c \in \{s, w, n\}$,

$$\mathbf{g}_c = \text{Softmax}(\mathbf{u}_c / T_g), \tag{15}$$

$$\boldsymbol{\ell}_c = \sum_{j \in \mathcal{A}} g_{c,j} E_j(z_c) + g_{c,0} E_0(z_c), \tag{16}$$

$$\boldsymbol{\ell} = p_s \boldsymbol{\ell}_s + p_w \boldsymbol{\ell}_w + p_n \boldsymbol{\ell}_n, \qquad \widehat{\mathbf{y}} = \text{Softmax}(\boldsymbol{\ell}). \tag{17}$$

$T_g > 0$ is the expert-gating temperature. Importantly, the same permission variable governs both branch contribution and external-information access, so routing is not merely a mixture weight over otherwise identical experts.

3.6 Training Objective

For a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$, authenticity is supervised by two-class cross-entropy,

$$\mathcal{L}_{\text{cls}} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{b \in \{0,1\}} y_{i,b} \log \widehat{y}_{i,b}. \tag{18}$$

Fake samples additionally have non-exclusive mechanism annotations $\mathbf{a}_i \in \{0,1\}^N$, used only for training-time auxiliary supervision and analysis. Because these annotations need not exhaust all plausible mechanisms, we form a smoothed mechanism-affinity target

$$\widetilde{\mathbf{a}}_i = \begin{cases} \mathbf{1}, & y_i = 0, \\ \mathbf{a}_i + \rho(\mathbf{1} - \mathbf{a}_i), & y_i = 1, \end{cases} \tag{19}$$

where $\rho \in [0, 1]$ assigns weak affinity to unmarked fake mechanisms. For authentic samples, the dense target avoids selecting a particular manipulation mechanism; for fake samples, smoothing avoids treating every unannotated mechanism as a certain negative. The same vectors are stored with training precedents and used by Strong and Weak. In the target-construction ablation *Authentic target* $\mathbf{0}$, only the authentic-sample target is replaced by $\mathbf{0}$; fake-sample targets remain unchanged.

Mechanism supervision follows the same route mixture as authenticity prediction. Let $\mathbf{s}_{i,c} = h_{\text{mech}}(z_{i,c})$ and $\overline{\mathbf{s}}_i = \sum_{c \in \{s,w,n\}} p_{i,c} \mathbf{s}_{i,c}$. With per-mechanism positive weights $\boldsymbol{\xi}$ estimated from the training targets, define

$$\ell_i^{\text{mech}} = -\frac{1}{N} \sum_{j=1}^{N} \left[\xi_j \widetilde{a}_{i,j} \log \sigma(\overline{s}_{i,j}) + (1 - \widetilde{a}_{i,j}) \log(1 - \sigma(\overline{s}_{i,j})) \right],$$

$$\mathcal{L}_{\text{mech}} = \omega_r \frac{1}{|\mathcal{D}_r|} \sum_{i \in \mathcal{D}_r} \ell_i^{\text{mech}} + \omega_f \frac{1}{|\mathcal{D}_f|} \sum_{i \in \mathcal{D}_f} \ell_i^{\text{mech}}, \tag{20}$$

where $\mathcal{D}_r$ and $\mathcal{D}_f$ are authentic/fake subsets and ω_r, ω_f prevent either class from dominating the auxiliary loss.

To discourage expert collapse, we average gates across samples and permission branches, $\overline{\mathbf{g}} = \frac{1}{3M} \sum_i \sum_{c \in \{s,w,n\}} \mathbf{g}_{i,c}$, and penalize deviation from uniform expert utilization:

$$\mathcal{L}_{\text{bal}} = \left\| \overline{\mathbf{g}} - \frac{1}{N+1} \mathbf{1}_{N+1} \right\|_2^2, \qquad \mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{mech}} \mathcal{L}_{\text{mech}} + \lambda_{\text{bal}} \mathcal{L}_{\text{bal}}. \tag{21}$$

Additional coefficient, architecture, optimization, and ablation-implementation details are given in Appendix C.3.

4 Experiments

4.1 Experimental Setup

Datasets and protocol. FakeSV contains 3,624 Chinese Douyin/Kuaishou videos (1,810 fake and 1,814 real; 39.88 s average), and FakeTT contains 1,991 English TikTok videos associated with

Table 1: Compact comparison on FakeSV and FakeTT (%). T/E denote temporal/event splits. We report ACC and Macro-F1; complete M-P/M-R results and provenance are in Appendix D.1 and Appendix D.14. †: unified baseline values reported by TrueLens; ‡: original paper; ∗: matched reproduction; –: unavailable.

	FakeSV				FakeTT			
Method	T-ACC	T-M-F1	E-ACC	E-M-F1	T-ACC	T-M-F1	E-ACC	E-M-F1
Unimodal								
BERT†	77.43	77.40	75.83	75.42	70.79	68.96	68.67	67.56
ViT†	71.34	71.04	72.94	72.90	64.32	64.15	70.24	68.15
Wav2Vec†	69.58	68.95	62.94	62.90	67.54	65.79	68.93	65.41
Conventional multimodal								
TikTec†	72.21	72.03	60.59	60.01	66.23	66.02	60.69	60.32
FANVM†	78.89	78.41	76.54	76.49	73.24	72.29	75.43	73.42
CAFE†	71.99	71.01	75.23	75.01	70.81	71.19	72.79	71.01
SV-FEND†	80.83	80.34	78.81	78.80	77.14	75.68	76.68	75.42
FakingRecipe†	85.01	85.00	80.53	80.52	80.01	78.23	75.25	73.90
ExMRD†	86.34	85.99	79.90	79.06	85.01	84.12	**78.14**	74.90
TGFC-SVFN‡	91.99	91.50	81.73	81.66	–	–	–	–
Multi-expert detector								
AMFNVD‡	87.26	87.00	–	–	84.28	82.21	–	–
Retrieval-augmented								
SV-FEND-SNEED‡	82.66	81.67	76.48	76.06	–	–	–	–
REAL†	77.49	77.44	73.00	73.04	72.24	70.10	69.78	66.82
TrueLens∗	93.47	93.35	88.44	88.43	85.55	84.41	78.12	76.86
PRME (ours)	**94.21**	**94.10**	**90.97**	**90.97**	**87.63**	**86.45**	77.87	**76.94**

Snopes events (1,172 fake and 819 real; 47.69 s average) (Qi et al., 2023a; Bu et al., 2024). Following the original study, each dataset is segmented in two ways: by event and by time. Each model and precedent bank is trained independently within its dataset. Mechanism labels are used only for training-time auxiliary supervision and analysis.

Baselines and implementation. We compare with unimodal, conventional multimodal, retrieval-augmented, and multi-expert detectors. We retain TrueLens feature pipeline: 16 visual/audio segments, CN-CLIP-L14 on FakeSV, LongCLIP-L14 on FakeTT, Wav2Vec audio, Whisper transcripts, and MLLM theme/storyline text. Hidden size is 256, batch size 32, learning rate 5×10^{-5}, weight decay 5×10^{-4}, dropout 0.30, and training lasts at most 20 epochs with AdamW and early stopping. Defaults are $k = 10$, $(\tau_s, \tau_w) = (0.82, 0.65)$, $T_r = 0.10$, and $T_g = 2.0$. We report ACC, Macro-F1 (M-F1), Macro-Precision (M-P), and Macro-Recall (M-R).

4.2 Performance Comparison and Analysis

Table 1 shows a split-dependent performance gap across both datasets. Relative to reproduced TrueLens, PRME gains +0.74/+0.75 ACC/M-F1 points on FakeSV temporal and +2.53/+2.54 on FakeSV event; on FakeTT temporal the gains are +2.08/+2.04. Event isolation is intentionally harder because same-event overlap is removed, reducing opportunities for reliable factual transfer and lowering absolute scores. On FakeTT event, the two methods are effectively tied (ACC -0.25, M-F1 $+0.08$); paired-bootstrap intervals span zero and event-fold standard deviations are about 4–6 points (Appendix D.12). We therefore interpret the strong temporal gains and FakeSV event gains as the stable improvements, while avoiding a claim of a reliable FakeTT event advantage.

Component ablations. Table 2 summarizes representative component ablations on the temporal splits of both datasets.

Table 2: Representative component ablations on the temporal splits of FakeSV and FakeTT. Macro-F1 mean $\pm$ standard deviation (%).

Variant	FakeSV Temporal	FakeTT Temporal
Evidence-permission routing		
Uniform branch weights	93.859 ± 0.466	85.466 ± 0.697
Hard relation routing	93.582 ± 0.619	84.577 ± 1.545
Random hard routing	93.445 ± 0.452	84.752 ± 1.895
Strong-only branch	93.802 ± 0.456	83.876 ± 1.029
Weak-only branch	92.900 ± 0.294	83.804 ± 0.657
No-evidence-only branch	93.383 ± 0.541	84.372 ± 1.065
$\beta_e = 0$	93.408 ± 0.439	85.616 ± 0.978
Mechanism-aware expert gating		
No Strong feature transfer	93.133 ± 0.236	84.653 ± 1.332
No Strong mechanism prior	93.640 ± 0.487	84.933 ± 1.261
No Weak mechanism prior	93.895 ± 0.488	84.818 ± 0.939
No external mechanism priors	93.615 ± 0.191	83.275 ± 1.014
Mechanism-agnostic MoE	93.349 ± 0.165	84.674 ± 1.257
Authentic target $\mathbf{0}$	93.490 ± 0.205	84.570 ± 0.710
General-expert fallback		
No general expert	94.018 ± 0.235	84.353 ± 0.887
Capacity controls		
Parameter-matched single expert	93.438 ± 0.181	84.988 ± 1.341
Parameter-matched deep MLP	93.358 ± 0.447	83.923 ± 0.872
Parameter-matched wide MLP	92.864 ± 0.618	84.347 ± 1.896
Training loss		
$\mathcal{L}_{cls}$ only (no mechanism loss)	93.721 ± 0.746	85.020 ± 1.314
No expert-balance loss	93.795 ± 0.360	84.789 ± 1.353
Full PRME	$\mathbf{94.103 \pm 0.513}$	$\mathbf{86.450 \pm 1.113}$

(1) Evidence-permission routing. Full PRME outperforms uniform weighting, hard relation routing, and random hard routing by 0.244/0.984, 0.521/1.873, and 0.658/1.698 Macro-F1 points on FakeSV/FakeTT, respectively. All single-branch variants are lower, with Weak-only dropping 1.203/2.646 points. Setting the event-consistency coefficient in Eq. equation 6 to $\beta_e = 0$ removes the same-event term while retaining text/visual support and lowers Macro-F1 by 0.695/0.834 points. These results support sample-dependent soft routing and complementary branch fusion.

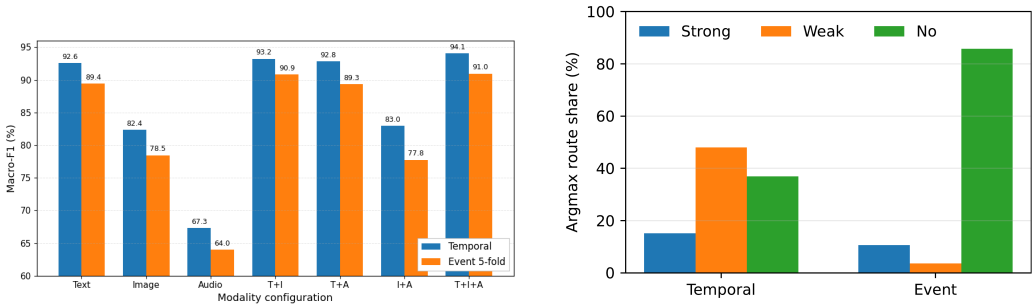
(2) Mechanism-aware expert gating. Removing Strong feature transfer, all external priors, and mechanism-aware gating lowers Macro-F1 by 0.970/1.797, 0.489/3.175, and 0.754/1.776 points on FakeSV/FakeTT, respectively. The *Authentic target* $\mathbf{0}$ ablation replaces the default all-one target for authentic samples in Eq. equation 19 with an all-zero vector, leaving fake-sample targets unchanged; it lowers Macro-F1 by 0.613/1.880 points. Together, these ablations support mechanism-specific gating and the default mechanism-target construction.

(3) Fallback, capacity, and training controls. Removing the general expert changes Macro-F1 by $-0.086/-2.097$ points on FakeSV/FakeTT. Parameter-matched single/deep/wide alternatives are all below Full PRME, with drops of 0.665/1.462, 0.745/2.527, and 1.239/2.103 points, respectively, showing that the gains are not explained by parameter count, depth, or width alone. Removing mechanism supervision or expert balancing also reduces temporal Macro-F1 by 0.382/1.430 and 0.308/1.661 points.

(4) Modal ablation. As shown in Figure 3(a), text-only reaches 92.58/89.42 M-F1 on temporal/event splits. Text+image is strongest among two-modality settings and is only 0.11 points below full PRME on the event split; audio contributes more clearly temporally than under event isolation.

4.3 Evidence Utilization Analysis

Figure 3(b) shows that evidence utilization adapts to retrieval conditions. Under the temporal split, Strong and Weak branches dominate, whereas under event isolation 85.75% of samples are routed to



(a) Modality ablation **(b)** Route distribution

Figure 3: FakeSV analyses. (a) End-to-end modality ablation: text is the strongest single modality and text+image is the strongest two-modality setting. (b) Argmax route distribution: event isolation shifts most samples to the internal No route.



(a) Routing temperature (b) Retrieval size

Figure 4: FakeSV sensitivity to routing temperature and retrieval size. Validation selects $T_r = 0.10$ and $k = 10$; the event gain at $k = 20$ is within fold variation.

the No branch, indicating that PRME automatically suppresses unreliable external evidence when transferable references are unavailable.

Relation-stratified results further localize the gains: relative to TrueLens, PRME changes Accuracy by -0.16, $+1.22$, and $+5.42$ points for same-event, similar-event, and same-topic queries, respectively. The detailed paired intervals and significance tests are reported in Appendix D.8; we treat this as evidence localization rather than a monotonicity claim because the similar-event interval crosses zero and the same-topic subset is small.

Sensitivity. Figure 4 reports routing-temperature and retrieval-size sensitivity on FakeSV. The validation-selected defaults $T_r = 0.10$ and $k = 10$ remain competitive on both splits; the small event increase at $k = 20$ is below the across-fold standard deviation, so we retain the unified defaults.

5 Conclusion

We presented Permission-Routed Mechanism Experts (PRME) for short-video misinformation detection. PRME jointly improves internal and external evidence modeling by combining mechanism-specific reasoning with permission-controlled precedent utilization. Mechanism experts enable specialized reasoning over diverse manipulation patterns from multimodal content, while the hierarchical Strong, Weak, and No router determines whether retrieved precedents contribute factual representations, mechanism priors, or no external information. Experiments on FakeSV and FakeTT demonstrate the effectiveness of the proposed framework across temporal and event settings. These results suggest that reliable short-video misinformation detection benefits from both specialized modeling of manipulation mechanisms and selective transfer of external evidence, rather than treating retrieved evidence as universally reusable.

AI USE STATEMENT

Generative AI tools were used to assist the sample-wise manipulation-type annotation described in Appendix B, and for language editing, formatting assistance, and manuscript layout refinement. The annotation procedure, evidence-matching constraints, and prompts are documented in the appendix. The authors take responsibility for the final manuscript, reported experiments, and all AI-assisted text or artifacts.

REPRODUCIBILITY STATEMENT

The main paper specifies the model architecture, routing equations, training objectives, datasets, evaluation protocol, and principal hyperparameters. The appendix provides the annotation protocol, implementation details, complete results, robustness analyses, and additional ablations needed to reproduce and audit the reported findings.

REFERENCES

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10: 746–763, 2022. doi: 10.1162/tacl_a_00486.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020.

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 *Proceedings of the ACM International Conference on Multimedia*, pp. 1351–1360, 2024.

Junyi Chen, Qian Liu, Jing Sun, and Yi Zhang. TrueLens: Video fake news detection with dual level evidence gathering and consolidation. In *Proceedings of the ACM Web Conference*, pp. 7178–7188, 2026. doi: 10.1145/3774904.3792362.

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

Zeyun Cheng, Liyuan Zhang, Yan Yang, and Yong Liu. AMFNVD: Addressing two kinds of discrepancy problems for fake news detection on short video platforms. In *Neural Information Processing*, volume 2757 of *Communications in Computer and Information Science*, pp. 374–387. Springer, 2026. doi: 10.1007/978-981-95-4100-3_27.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Guoxuan Ding, Yuqing Li, Ziyan Zhou, Zheng Lin, Daren Zha, and Jiangnan Li. ExDR: Explanation-driven dynamic retrieval enhancement for multimodal fake news detection. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 290–300, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Richang Hong, Jieming Lang, Jie Xu, Zhiyong Cheng, Tao Zhong, and Fei Zhou. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM Web Conference*, pp. 4684–4698, 2025.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Nuo Li, Yang Xiong, Chenglizhao Liu, Jie Wen, and Chao Huang. Detecting fake news in short videos through multi-view aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 632–640, 2026a. doi: 10.1609/aaai.v40i1.37028.

Yiheng Li, Weihai Lu, Hanyi Yu, and Yue Wang. Retrieval-augmented multimodal model for fake news detection. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 969–980, 2026b. doi: 10.1145/3805712.3809605.

Yili Li, Jian Lang, Rongpei Hong, Qing Chen, Zhangtao Cheng, Jia Chen, Ting Zhong, and Fan Zhou. REAL: Retrieval-augmented prototype alignment for improved fake news video detection. In *2025 IEEE International Conference on Multimedia and Expo*, 2025. doi: 10.1109/ICME59968.2025.11209008.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. MDFEND: Multi-domain fake news detection. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 3343–3347, 2021. doi: 10.1145/3459637.3482139.

Zhuangzhuang Pan, Yan Xia, Vimala Balakrishnan, Muhammad Zaiamri Zainal Abidin, and Siti Soraya Abdul Rahman. DERM: Dynamic evidence reliability modeling for multimodal rumor detection. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 107:1–107:7, 2025. doi: 10.1145/3743093.3771053.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Jun 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, Yuyan Zhao, Yang 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.

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. doi: 10.1109/BigData52589.2021.9671928.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. SpotFake: A multi-modal framework for fake news detection. In *IEEE International Conference on Multimedia Big Data*, pp. 39–47, 2019.

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, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.257.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 849–857, 2018.

Table B.1: Dataset statistics used in this work.

Dataset	Total	Fake	Real	Avg. sec.	Language/platform
FakeSV	3,624	1,810	1,814	39.88	Chinese; Douyin/Kuaishou
FakeTT	1,991	1,172	819	47.69	English; TikTok/Snopes

Guangyu Wei, Ke Han, Yueming Lyu, Yu Luo, Yue Jiang, Caifeng Shan, and Nicu Sebe. Towards real-time fake news detection under evidence scarcity. *arXiv preprint arXiv:2510.11277*, 2025.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5384–5392, 2023. doi: 10.1609/aaai.v37i4.25670.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE: Similarity-aware multi-modal fake news detection. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 354–367, 2020.

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 Additional Method Notes

Section 3 gives the complete architecture and training-objective equations. To avoid duplicating the main paper, we do not restate those equations here. The remaining appendices provide the dataset annotation protocol, implementation coefficients and optimization settings, ablation definitions, and complete experimental results; the implementation-specific settings are collected in Appendix C.3.

B Dataset Annotation and Empirical Findings

B.1 Dataset Overview

Table B.1 summarizes the data. FakeSV contains 3,624 Chinese short videos collected from Douyin and Kuaishou, including 1,810 fake and 1,814 real samples, with an average duration of 39.88 seconds. FakeTT contains 1,991 English TikTok videos associated with Snopes events, including 1,172 fake and 819 real samples, with an average duration of 47.69 seconds. Following the original dataset protocols, both datasets are evaluated with temporal and event-based partitions Qi et al. (2023a); Bu et al. (2024). The mechanism labels below are used only for training-time auxiliary supervision and analysis, not at test time.

B.2 Coarse Event and Domain Annotation

We augmented the original event identifiers with two coarser labels. The `domain_label` records one of nine broad domains (disaster, society, health, culture, politics, science, education, finance, and military), while `event_keyword` gives a location- and time-independent event summary, such as "train-seat dispute," "firefighter death," or "traffic accident." FakeSV's original `keywords` field is treated as the fine-grained event identifier; FakeTT uses its original `event` field for the same role. Table B.2 summarizes the hierarchy `domain_label` → `event_keyword` → `keywords/event` and the corresponding FakeSV label cardinalities.

Prompt for FakeSV Add two labels to every sample and output a new file. First, assign a `domain_label`, using categories such as disaster, society, health, culture, politics, science, education, finance, and military. Second, assign an `event_keyword`. This keyword should summarize the general event without a specific time or location; examples include "train-seat dispute," "firefighter death," and "traffic accident." The keyword will be used to retrieve similar events, so it should be coarser than the existing `keywords` field.

Table B.2: FakeSV label hierarchy after augmentation.

Field	Full	Train	Test	Examples
domain_label	9	9	9	disaster, society, health
event_keyword	361	319	139	pandemic, factory fire, typhoon
keywords	614	542	206	Pope Francis dancing merengue

Prompt for FakeTT Add two labels to every sample and output a new file. Assign a `domain_label` from broad domains such as disaster, society, health, culture, politics, science, education, finance, and military. Assign an `event_keyword` that summarizes the general event without a specific time or location. The keyword will be used to retrieve similar events and should therefore be coarser than FakeTT's existing fine-grained `event` field.

B.3 Manipulation-Type Annotation

We used a GPT-5.6 Thinking configuration to annotate fake samples. The same sample-wise evidence-verification template was used for FakeSV and FakeTT. The model was instructed not to cluster samples before verification and not to infer labels from general knowledge. It first extracted the event claim, constructed multilingual search queries, searched each sample independently, and retained an evidence match only when at least three of the following attributes agreed: entity, event, location, time, and visual claim.

Manipulation-label prompt **Role.** You are a misinformation-detection expert. Process the uploaded JSONL file.

Step 1. Select all samples whose `annotation` is fake.

Step 2. Process every fake sample independently before any clustering. Use `keywords` only as auxiliary information and rely mainly on `title`, `description`, and `temporal_evolution`. Produce `sample_id`, `event_summary`, `key_entities`, `key_action`, `claimed_time`, `claimed_location`, `visual_claim`, and `search_queries`.

Step 3. Generate at least five Chinese search queries: the original title plus year; the core description plus year; the most specific sentence from `temporal_evolution` plus year; entity + action + location + a Chinese equivalent of "debunk"; and entity + action + location + a Chinese equivalent of "rumor," "false," or "online claim." For overseas events, also generate at least three English queries using "fact check," "false," and "misleading" or "hoax."

Step 4. Search every sample separately for at least two rounds: first with the original wording and then with a rewritten entity–action–location– year query. If both fail, use the most specific sentence from `temporal_evolution`. Use publication time only to identify the year and allow the preceding and following year. Prioritize official notices, debunking platforms, fact-checking sites, and verification reports from major media. Do not judge from common knowledge.

Step 5. Accept an evidence match only when at least three of entity, event, location, time, and visual claim agree. Mark partial matches as weak evidence rather than forcing a conclusion.

Step 6. Add multi-label `fake_type` annotations selected from: temporal mismatch, spatial mismatch, entity mismatch, causal disorder, insufficient evidence, synthesis, fabricated or exaggerated facts and commonsense violation. Also add `fake_explanation`, `evidence_sources`, `evidence_strength`, and `attempted_queries`.

Step 7. Output (i) `fake_samples_checked.jsonl`, (ii) `fake_event_clusters_checked.csv` after sample-wise verification, and (iii) `search_log.csv`. Do not merge samples merely because their keywords match, do not transfer one event cluster's conclusion to a different event, and preserve unverified samples as "insufficient evidence."

Table B.3 lists the eight reported FakeSV base labels. Because the annotation is multi-label, the total number of label occurrences exceeds the number of fake samples.

B.4 Detailed Analysis of the Introductory Empirical Finding

We define four mutually exclusive query–precedent relation levels. A query has a *same-event* precedent when the fine-grained event identifier matches; a *similar-event* precedent when `event_keyword` matches but the fine-grained event differs; a *same-topic* precedent when only `domain_label` matches; and an *unrelated* precedent when none of these labels match. Each query is compared with all other samples in the corresponding analysis set.

Table B.3: Reported FakeSV manipulation-label counts.

Label	Full	Train	Test
Insufficient evidence	190	133	17
Temporal mismatch	341	223	80
Location mismatch	456	313	86
Subject mismatch	327	210	62
Causal disorder	192	143	35
Post-processing/synthesis	266	188	33
Commonsense violation	168	100	49
Fabricated/exaggerated facts	1,493	1,004	271
Total label occurrences	3,433	2,314	633
Fake samples	1,810	1,233	304

Table B.4: Manipulation-type consistency by relation level on FakeSV.

Relation	n	Share	Avg. refs	Dominant	Entropy	Ref–ref J	Target–ref J	Exact agreement
All 3,624 samples; authentic samples are treated as an additional "real" type								
Same event	3531	97.43%	8.32	0.939	0.178	0.894	0.889	0.888
Similar event	49	1.35%	134.96	0.610	1.871	0.499	0.299	0.214
Same topic	44	1.21%	883.64	0.440	3.000	0.322	0.236	0.128
Unrelated	0	0.00%	–	–	–	–	–	–
Only 1,810 fake samples								
Same event	1680	92.82%	6.94	0.986	0.039	0.976	0.975	**0.973**
Similar event	71	3.92%	74.80	0.420	2.491	0.461	0.327	**0.117**
Same topic	59	3.26%	474.03	0.247	3.628	0.364	0.321	**0.106**
Unrelated	0	0.00%	–	–	–	–	–	–

For each reference set, we report the dominant exact-label-combination ratio, the entropy of exact manipulation-label combinations, the mean pairwise multi-label Jaccard similarity within the reference set, the target–reference Jaccard similarity, and the exact-type agreement produced by the reference set. Groups with fewer than two samples are excluded from consistency aggregation. For two multi-label sets A and B, Jaccard similarity is

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{B.1}$$

A high dominant ratio and Jaccard score together with low entropy indicate that a relation level provides a stable manipulation prior.

Tables B.4 and B.5 explain the observation summarized in the introduction. Same-event precedents have near-deterministic mechanism consistency: their exact agreement is 0.973, target–reference Jaccard is 0.975, and combination entropy is only 0.039. Similar-event precedents retain partial mechanism overlap but do not reliably preserve the complete manipulation pattern: exact agreement falls to 0.117 and entropy rises to 2.491. Same-topic precedents are still less specific, with 0.106 exact agreement and entropy 3.628. Thus, semantic or topical proximity is not equivalent to either factual transferability or exact mechanism transferability.

The temporal test-set coverage gives a complementary, less restrictive view. Although a similar-event or same-topic reference set often contains at least one overlapping label, the average target–reference Jaccard remains low. This means that **weak precedents can provide a coarse candidate mechanism** but should not overwrite query-specific entities, times, locations, or event facts.

Table B.6 shows that reproduced TrueLens depends strongly on event relation: Accuracy is 96.30% for same-event queries but drops to 87.30% and 87.80% for similar-event and same-topic queries, respectively.

Accordingly, Table B.7 maps same-event evidence to factual transfer and mechanism guidance, similar-event evidence to a weak mechanism prior, same-topic evidence to broad background only, and unrelated or unsupported cases to an internal-evidence fallback. PRME does not hard-code these symbolic labels at inference time; they motivate the Strong, Weak, and No permissions, while the router assigns soft probabilities from retrieval consistency scores.

Table B.5: Reference availability on the FakeSV temporal test set. "Exact available" means that at least one reference has the target's exact label combination; "any label" means that at least one mechanism label overlaps.

Relation	n	Share	Exact available	Any label available	Mean target–ref J
All 542 test samples					
Same event	368	67.90%	359 (97.55%)	359 (97.55%)	0.9183
Similar event	115	21.22%	62 (53.91%)	108 (93.91%)	0.2790
Same topic	59	10.89%	49 (83.05%)	59 (100.00%)	0.2810
Unrelated	0	0.00%	0	0	–
Only 304 fake test samples					
Same event	190	62.50%	184 (96.84%)	184 (96.84%)	0.9314
Similar event	76	25.00%	26 (34.21%)	72 (94.74%)	0.1756
Same topic	38	12.50%	28 (73.68%)	38 (100.00%)	0.1661
Unrelated	0	0.00%	0	0	–

Table B.6: Relation-stratified performance of reproduced TrueLens on the FakeSV temporal test set.

Relation	n	ACC	M-F1	M-P	M-R
Same event	368	0.9630	0.9643	0.9616	0.9674
Similar event	115	0.8730	0.9041	0.9014	0.9079
Same topic	59	0.8780	0.9093	0.8758	0.9474
Overall	542	0.9347	0.9422	0.9351	0.9500

Table B.7: Empirically motivated evidence permissions.

Evidence relation	Permitted use
Same event	Strong: factual interaction and mechanism prior
Similar event	Weak: mechanism-level guidance without fact injection
Same topic	Weak or No: background only, depending on support
Unrelated/unsupported	No: internal multimodal evidence only

C Experimental Design Details

C.1 Baselines and Result Provenance

We compare unimodal, conventional multimodal, retrieval-augmented, and multi-expert detectors. Table C.1 summarizes the role of each baseline and the provenance of the reported numbers. Each citation in the Method column refers to the original paper that introduced the corresponding method or backbone.

C.2 Training Time and GPU Memory Consumption

Training was conducted on a single NVIDIA L40S GPU with 48 GB of memory, taking 5.02 minutes for the training loop and 5.19 minutes end-to-end, with a peak allocated GPU memory usage of 19.97 GB and a peak reserved memory usage of 35.60 GB.

C.3 Reproducibility and Evaluation Protocol

We retain TrueLens Chen et al. (2026) feature pipeline: each video is uniformly represented by 16 visual and audio segments. Text and visual inputs are encoded by CN-CLIP-L14 on FakeSV and LongCLIP-L14 on FakeTT; audio features are extracted by Wav2Vec and transcripts by Whisper. The hidden size is 256, batch size is 32, learning rate is 5×10^{-5}, weight decay is 5×10^{-4}, dropout is 0.30, and the maximum number of epochs is 20. We optimize with AdamW and stop after five validation epochs without improvement.

The default model uses nine experts (eight mechanism experts and one general expert), three-layer MLP expert heads, $k = 10$ retrieved precedents, $(\tau_s, \tau_w) = (0.82, 0.65)$, routing temperature $T_r = 0.10$, and expert-gate temperature $T_g = 2.0$. The mechanism-loss and balance-loss weights are 0.1 and 0.01. The branch coefficients are $\lambda_{seen} = 0.5$, $\lambda_{weak} = 1.0$, $\lambda_{no} = 0.5$, and the general-expert fallback bias is $\lambda_{common} = 0.5$.

Table C.1: Concise baseline descriptions and numerical sources.

Method	Main idea	Result source
BERT / ViT / Wav2Vec Devlin et al. (2019); Dosovitskiy et al. (2021); Baevski et al. (2020)	Text-only, visual-only, and audio-only reference encoders.	Unified TrueLens protocol
TikTec Shang et al. (2021)	Detects COVID-19 misinformation in TikTok videos by jointly modeling caption-guided visual, audio, and textual cues.	Unified TrueLens protocol
FANVM Choi & Ko (2021)	Uses topic modeling and adversarial learning to obtain topic-agnostic fake-news video representations.	Unified TrueLens protocol
CAFE Chen et al. (2022)	Learns cross-modal ambiguity for multimodal fake-news detection.	Unified TrueLens protocol
SV-FEND Qi et al. (2023a)	Integrates short-video multimodal content and rich social context.	Unified TrueLens protocol
FakingRecipe Bu et al. (2024)	Models creative-process patterns used to construct short-video misinformation.	Unified TrueLens protocol
ExMRD Hong et al. (2025)	Performs explainable micro-video rumor detection through chain-of-thought clue reasoning.	Unified TrueLens protocol
TGFC-SVFN Zong et al. (2025)	Uses text-guided fine-grained counterfactual inference.	Original paper
SV-FEND-SNEED Qi et al. (2023a;b)	Augments SV-FEND with event-neighbor retrieval based on NEED.	Original paper
REAL Li et al. (2025)	Retrieves relevant real and fake videos and aligns the target with class-specific prototypes.	Unified TrueLens protocol
TrueLens Chen et al. (2026)	Gathers real and fake precedents and consolidates evidence at two levels.	Matched reproduction in our tables
AMFNVD Cheng et al. (2026)	Uses an adaptive multi-gate mixture of experts and multi-branch joint loss to address feature-distribution and modality-contribution discrepancies.	Original paper

PCS/TSS coefficients. Default runs read the precomputed PCS/TSS values from the frozen training-only retrieval artifact. When routing scores must be recomputed for ablations that alter event-ID usage, Eq. equation 5 uses $(\gamma_t, \gamma_v, \gamma_a) = (0.30, 0.58, 0.12)$, while Eq. equation 6 uses text/visual coefficients $(\beta_t, \beta_v) = (0.70, 0.10)$ and the event coefficient β_e for the same-event indicator. The $\beta_e = 0$ ablation sets only that event coefficient to zero; when event IDs are disabled, the same-event indicator is identically zero. In all other cases, the frozen TSS is used verbatim.

All temporal results are averaged over seeds 2023–2027. Event results use five event-isolated folds and are paired by fold with the reproduced TrueLens baseline. FakeSV and FakeTT are trained independently. The precedent bank is rebuilt for every seed or fold using training samples only; query self-matches are removed, and no test sample or test-derived representation is inserted. Every FakeSV temporal run evaluates the same 542 test videos; the event-fold test sizes are 717, 725, 726, 727, and 729.

We report Accuracy (ACC), Macro-Precision (M-P), Macro-Recall (M-R), and Macro-F1 (M-F1). For class set $\mathcal{Y} = \{0, 1\}$,

$$\text{M-F1} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{F1}_c, \tag{C.1}$$

with M-P and M-R defined analogously. For paired statistical analysis, we compute PRME minus TrueLens. The 95% bootstrap interval exhaustively enumerates all $5^5 = 3{,}125$ paired resamples. We additionally report paired two-sided t-tests, Holm family-wise correction, Benjamini–Hochberg false-discovery-rate correction, and an exact two-sided sign-flip test. With five pairs, the minimum attainable exact two-sided p-value is 0.0625.

C.4 Implementation of Ablation Groups

Unless explicitly changed, the encoders, optimizer, 20-epoch schedule, losses, data split, and retrieval bank are fixed. The default route uses $k = 10$, $(\tau_s, \tau_w) = (0.82, 0.65)$, $T_r = 0.10$, mechanism-loss weight 0.1, balance-loss weight 0.01, and nine experts.

Permission-router controls. (1) **Uniform branch weights**: fixes $(p_s, p_w, p_n) = (1/3, 1/3, 1/3)$; (2) **Rolled weights**: cyclically shifts the learned route weights by one position in each batch while preserving their overall distribution; (3) **Strong branch only**, (4) **Weak branch only**, and (5) **No branch only**: use the corresponding one-hot routes; (6) **Hard relation routing**: takes the argmax of

the original route; (7) **Random soft routing**: uses a deterministic video-ID/seed hash to generate a Dirichlet$(1, 1, 1)$ weight vector; and (8) **Random hard routing**: maps each video deterministically to one branch. These controls leave the branch representations and expert pool unchanged.

Score and target ablations. $\beta_e = 0$ removes the same-event term from TSS while retaining its text/visual terms. **Authentic target 0** replaces the default all-one mechanism-affinity target for authentic samples with an all-zero vector; fake-sample targets and all other components are unchanged.

Alternative transfer and aggregation controls. (1) **Weak-Fact** tests whether Weak evidence should directly inherit the Strong branch's fact-enhanced representation. It changes only the Weak branch input, internal mechanism prediction, and expert input from z_{base} to z_{seen}; the learned route, Weak Top-k mechanism prior, Strong/No branches, expert pool, and losses remain fixed. (2) **Continuous Reliability MoE** only has one branch. It tests whether three condition-specific computations can be replaced by one continuously interpolated computation. It forms

$$z_{cont} = p_s z_{seen} + (1 - p_s) z_{base}, \tag{C.2}$$

$$r_{cont} = p_s \lambda_{seen} r_s + p_w \lambda_{weak} r_w. \tag{C.3}$$

and only invokes the nine-expert MoE once. In contrast, Full PRME computes Strong, Weak, and No branch outputs independently and combines their logits at the end.

Modality controls. (1) **Text only**, (2) **Image only**, (3) **Audio only**, (4) **Text + image**, (5) **Text + audio**, and (6) **Image + audio** retain the named modality subsets. For each variant, all other modalities are removed from both the model input and the retrieval/router computation, a train-only evidence bank is rebuilt, the text/image/audio retrieval weights are renormalized, and PCS/TSS are recomputed from the available modalities. The same-event term is used only when text remains; audio-only TSS falls back to audio similarity. Thus, these experiments measure the behavior of the complete system under missing modalities and may also alter the route distribution.

Expert-input control. (1) **Without Strong feature enhancement**: replaces the Top-1 cross-attended z_{seen} with z_{base} while keeping the Strong route and prior.

External-prior controls. (1) **Without Strong external prior**: zeros only the Strong mechanism prior; (2) **Without Weak external prior**: zeros only the Weak mechanism prior; and (3) **Without all external priors**: zeros both priors while preserving Strong feature enhancement.

Internal-mechanism-cue controls. (1) **Without Strong internal cue**: zeros only the Strong branch's mechanism logits; (2) **Without Weak internal cue**: zeros only the Weak branch's mechanism logits; (3) **Without No-route internal cue**: zeros only the No-route branch's mechanism logits; (4) **Without all internal mechanism cues**: zeros all three branch-specific mechanism logits.

Fallback control. (1) **Without fallback control**: removes the additional general-expert preference on the No route while retaining the same expert pool and all other gate terms.

Generic-gate control. (1) **Mechanism-agnostic MoE**: retains the nine-expert structure but removes the external priors and mechanism-label supervision, providing a capacity-matched test of semantic expert gating.

Loss controls. (1) **Without mechanism supervision**: removes the auxiliary mechanism-label loss; and (2) **Without expert-balance loss**: removes the expert-balance regularizer while leaving the architecture unchanged.

Expert-architecture controls. (1) **One-layer MLP**, (2) **Two-layer MLP**, (3) **Three-layer MLP (default)**, and (4) **Four-layer MLP**: change only the number of linear layers inside each expert; (5) **Transformer expert**: replaces each MLP expert with a Transformer-based expert while retaining the surrounding routing and gating modules.

Table D.1: Complete comparison on FakeSV under temporal and event splits (%). †: unified results reported by TrueLens; ‡: original paper; ∗: matched reproduction. FakeSV-VLM does not report event-split results.

	Temporal				Event			
Method	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
Unimodal								
BERT†	77.43	77.40	77.03	77.69	75.83	75.42	76.01	76.03
ViT†	71.34	71.04	70.57	71.71	72.94	72.90	72.97	72.01
Wav2Vec†	69.58	68.95	69.32	69.01	62.94	62.90	63.00	62.96
Conventional multimodal								
TikTec†	72.21	72.03	72.02	72.49	60.59	60.01	60.42	59.49
FANVM†	78.89	78.41	78.99	79.01	76.54	76.49	76.53	76.52
CAFE†	71.99	71.01	71.54	71.29	75.23	75.01	75.20	75.29
SV-FEND†	80.83	80.34	80.23	80.69	78.81	78.80	78.54	78.68
FakingRecipe†	85.01	85.00	84.99	85.05	80.53	80.52	80.54	80.61
ExMRD†	86.34	85.99	86.12	86.50	79.90	79.06	79.35	79.09
TGFC-SVFN‡	91.99	91.50	92.93	90.68	81.73	81.66	81.80	81.68
Retrieval-augmented								
SV-FEND-SNEED‡	82.66	81.67	84.65	81.03	76.48	76.06	78.22	76.46
REAL†	77.49	77.44	77.12	78.10	73.00	73.04	73.13	73.06
TrueLens (reproduced)∗	93.47	93.35	93.54	93.26	88.44	88.43	88.57	88.44
Multi-expert detector								
AMFNVD‡	87.26	87.00	–	–	–	–	–	–
PRME (ours)	**94.21**	**94.10**	**94.25**	**94.00**	**90.97**	**90.97**	**91.06**	**90.98**

Number-of-experts controls. (1) **One expert (same architecture)**: keeps one original three-layer expert; (2) **One expert (parameter matched)**: widens a shared three-layer MLP to approximate the parameter count of the full pool; (3) **Four merged-mechanism experts**: merges mechanism pairs; (4) **Eight experts, no general expert**: deletes the general expert; (5) **Nine experts (default)**: retains the complete pool; and (6) **Twelve experts**: adds three potential experts.

Module-composition controls. (1) **Internal classifier only**: removes all retrieval paths and the mechanism-aware expert pool; (2) **Three paths + single head**: retains the three evidence paths but replaces the expert pool with one classification head; (3) **Unified evidence + mechanism MoE**: unifies the evidence representation before mechanism-aware gating; and (4) **Three paths + mechanism MoE**: is the complete PRME composition.

Parameter-matched controls. (1) **Single internal branch (parameter matched)** retains only an internal-evidence classifier and increases its capacity to match the full model; (2) **Deep MLP without paths/experts** and (3) **Wide MLP without paths/experts** remove retrieval paths and experts and match the parameter scale with ordinary deep or wide classifiers, respectively.

D FakeSV Data Analysis

D.1 Complete FakeSV Main Results

Table 1 reports the compact ACC/Macro-F1 comparison for both datasets. Table D.1 restores Macro-Precision and Macro-Recall for FakeSV under both evaluation protocols.

Table D.2: Routing and modality ablations on FakeSV. Values are Macro-F1 mean ± standard deviation (%). Parentheses show the change from the full model.

Category	Variant	Temporal	Event
Routing	Uniform branch weights	93.859 ± 0.466 (-0.244)	90.666 ± 1.786 (-0.302)
	Rolled weights	93.433 ± 0.632 (-0.670)	90.804 ± 1.686 (-0.165)
	Strong branch only	93.802 ± 0.456 (-0.302)	90.386 ± 1.724 (-0.583)
	Weak branch only	92.900 ± 0.294 (-1.203)	90.525 ± 1.910 (-0.443)
	No branch only	93.383 ± 0.541 (-0.720)	90.852 ± 1.628 (-0.116)
	Weak-Fact	93.692 ± 0.207 (-0.411)	90.995 ± 1.434 $(+0.027)$
	Continuous Reliability MoE	93.778 ± 0.565 (-0.325)	90.743 ± 1.609 (-0.225)
	Hard relation routing	93.582 ± 0.619 (-0.521)	90.768 ± 1.685 (-0.200)
	Random soft routing	93.774 ± 0.281 (-0.329)	90.637 ± 1.449 (-0.331)
	Random hard routing	93.445 ± 0.452 (-0.658)	91.168 ± 1.458 $(+0.200)$
Modalities	Text only	92.580 ± 0.430 (-1.523)	89.424 ± 1.467 (-1.544)
	Image only	82.368 ± 0.623 (-11.735)	78.463 ± 2.076 (-12.505)
	Audio only	67.321 ± 0.641 (-26.782)	64.026 ± 1.737 (-26.942)
	Text + image	93.238 ± 0.389 (-0.865)	90.858 ± 1.494 (-0.110)
	Text + audio	92.834 ± 0.231 (-1.269)	89.330 ± 1.393 (-1.638)
	Image + audio	82.955 ± 0.496 (-11.148)	77.760 ± 2.073 (-13.208)
Full	PRME	$\mathbf{94.103 \pm 0.513}$	$\mathbf{90.968 \pm 1.641}$

Table D.3: Detailed results for Weak-Fact and Continuous Reliability MoE on FakeSV. Values are mean ± standard deviation (%). The final column is the mean predicted probability of the fake class.

Variant	Split	ACC	M-F1	M-P	M-R	$\mathbb{E}[p(y = \text{fake})]$
Weak-Fact	Temporal	93.801 ± 0.210	93.692 ± 0.207	93.834 ± 0.332	93.598 ± 0.222	0.5703 ± 0.0136
Continuous Reliability MoE	Temporal	93.875 ± 0.560	93.778 ± 0.565	93.818 ± 0.602	93.746 ± 0.540	0.5625 ± 0.0071
Weak-Fact	Event 5-fold	91.002 ± 1.437	90.995 ± 1.434	91.157 ± 1.501	91.010 ± 1.418	0.4843 ± 0.0265
Continuous Reliability MoE	Event 5-fold	90.754 ± 1.604	90.743 ± 1.609	90.931 ± 1.513	90.759 ± 1.590	0.4740 ± 0.0157

D.2 Complete FakeSV Ablations

Tables D.2, D.4, and D.5 provide the complete routing, modality, evidence-channel, gating, loss, architecture, and capacity controls summarized in the main paper.

D.3 Alternative Transfer and Single-Pass Aggregation

Table D.3 reports metrics not shown in the main paper for Weak-Fact and Continuous Reliability MoE. These controls separate two questions: whether Weak evidence should be promoted to fact-level feature transfer, and whether reliability can be compressed into one interpolated representation before expert reasoning.

As shown in Table D.3, on the temporal split, Weak-Fact and Continuous Reliability MoE are 0.411 and 0.325 M-F1 points below Full PRME. Weak evidence therefore appears more useful as a mechanism prior than as permission for direct fact-level feature transfer. Likewise, compressing reliability into one interpolated representation does not fully replace condition-specific expert computation followed by logit-level fusion. On the event split, Weak-Fact is 0.027 points above Full PRME, which is negligible relative to the 1.434-point across-fold standard deviation and the 3.65% Weak-route share. We consequently treat it as statistically indistinguishable rather than as an improvement. Continuous Reliability MoE remains 0.225 points below the full model.

D.4 Modality Removal Also Changes the Evidence Regime

The modality experiments rebuild retrieval and recompute PCS/TSS, so they are end-to-end missing-modality tests rather than classifier-only feature masks. Table D.6 makes the resulting route shift explicit for representative variants.

As shown in Table D.6, audio-only sends every sample to Strong, showing that PCS/TSS scales and the meaning of fixed thresholds change after removing modalities. The reported modality ablations therefore quantify the performance of the complete pipeline when modalities are unavailable; they do not isolate pure classifier-side information gain. A stricter attribution study would either freeze the full-model route weights or recalibrate thresholds inside the training data for every modality subset.

Table D.4: Expert inputs, gates, and losses on FakeSV. Values are Macro-F1 mean ± standard deviation (%).

Category	Variant	Temporal	Event
Expert input	Without Strong feature enhancement	93.133 ± 0.236	91.107 ± 1.743
External priors	Without Strong external prior	93.640 ± 0.487	90.745 ± 2.043
	Without Weak external prior	93.895 ± 0.488	90.991 ± 2.056
	Without all external priors	93.615 ± 0.191	90.878 ± 2.129
Internal mechanism cues	Without Strong internal cue	93.335 ± 0.411	90.720 ± 1.798
	Without Weak internal cue	93.454 ± 0.485	90.530 ± 1.309
	Without No-route internal cue	93.603 ± 0.573	90.774 ± 1.381
	Without all internal mechanism cues	93.553 ± 0.554	90.948 ± 1.496
Generic gate	Mechanism-agnostic MoE	93.349 ± 0.165	90.685 ± 1.876
Losses	Without mechanism supervision	93.721 ± 0.746	90.748 ± 1.406
	Without expert-balance loss	93.795 ± 0.360	90.968 ± 1.808
Full	PRME	$\mathbf{94.103 \pm 0.513}$	$\mathbf{90.968 \pm 1.641}$

Table D.5: Expert architecture, expert number, and capacity controls on FakeSV (Macro-F1, %).

Variant	Temporal	Event
Expert architecture		
One-layer MLP	93.292 ± 0.447	91.248 ± 1.093
Two-layer MLP	93.342 ± 0.350	90.859 ± 1.381
Three-layer MLP (default)	94.103 ± 0.513	90.968 ± 1.641
Four-layer MLP	93.771 ± 0.341	90.358 ± 1.768
Transformer expert	93.464 ± 0.306	90.445 ± 1.654
Number of experts		
One expert (same architecture)	93.436 ± 0.710	90.387 ± 1.554
One expert (parameter matched)	93.438 ± 0.181	90.245 ± 1.680
Four merged-mechanism experts	93.434 ± 0.288	90.360 ± 1.643
Eight experts, no general expert	94.018 ± 0.235	90.422 ± 2.040
Nine experts (default)	94.103 ± 0.513	90.968 ± 1.641
Twelve experts	93.453 ± 0.591	90.449 ± 1.257
Module composition		
Internal classifier only	93.136 ± 0.312	90.666 ± 1.403
Three paths + single head	93.647 ± 0.514	90.751 ± 1.488
Unified evidence + mechanism MoE	93.735 ± 0.159	90.448 ± 1.157
Three paths + mechanism MoE	94.103 ± 0.513	90.968 ± 1.641
Parameter-matched controls		
Deep MLP without paths/experts	93.358 ± 0.447	90.308 ± 1.714
Wide MLP without paths/experts	92.864 ± 0.618	90.129 ± 1.940

Table D.6: Argmax route distribution (%) after representative modality removal on FakeSV.

Split	Modalities	Strong	Weak	No
Temporal	Text	61.44	13.28	25.28
Temporal	Image	13.65	45.39	40.96
Temporal	Audio	100.00	0.00	0.00
Temporal	Text + image	13.47	51.66	34.87
Event	Text	42.62	0.00	57.38
Event	Image	7.78	49.90	42.32
Event	Audio	100.00	0.00	0.00
Event	Text + image	8.15	5.41	86.44

Table D.7: Argmax route proportions and mean route entropy for representative FakeSV routing controls.

Split	Routing	Strong	Weak	No	Entropy
Temporal	Full relation route	15.13	47.97	36.90	0.844
Temporal	Random soft	34.02	32.55	33.43	0.837
Temporal	Random hard	33.80	34.28	31.92	0.000
Event	Full relation route	10.60	3.65	85.75	0.950
Event	Random soft	33.67	32.50	33.83	0.834
Event	Random hard	33.11	33.89	33.00	0.000

Table D.8: Relation-stratified paired Accuracy on FakeSV temporal data. Differences are PRME minus reproduced TrueLens (percentage points).

Relation	n	Real/Fake	Mean difference [95% CI]	Paired t p	Sign-flip p
Same event	368	178/190	-0.16 $[-0.65, +0.33]$	.591	.750
Similar event	115	39/76	$+1.22$ $[-0.17, +2.61]$	.226	.250
Same topic	59	21/38	$+5.42$ $[+3.73, +7.12]$	.0054	.0625

D.5 Route-Distribution Diagnostics

Table D.7 reports route proportions and entropy for the learned and randomized controls.

Table D.7 shows that event isolation changes the evidence regime rather than merely lowering accuracy: the learned router assigns 85.75% of samples to the No route and only 3.65% to Weak. Random hard routing forces near-uniform branch usage and can achieve a similar mean M-F1, but Table D.11 shows that it is markedly less calibrated. This distinction motivates evaluating route behavior and probability quality together.

D.6 Relation-Stratified Paired Audit

The temporal relation file covers all 542 test videos without overlap: 368 same-event, 115 similar-event, and 59 same-topic cases. No unrelated case remains under the retrieval taxonomy, so a complete monotonic trend across four relation levels cannot be tested. Table D.8 reports paired Accuracy differences using the same five seeds for PRME and TrueLens.

As reported in Table D.8, the same-topic effect is positive for all five seeds, but the subset contains only 59 videos and the exact two-sided test cannot fall below 0.0625 with five pairs. The similar-event confidence interval crosses zero. We therefore present this analysis as evidence localization rather than a confirmatory monotonicity claim. It also resolves a metric inconsistency in an earlier draft: the reproducible paired Accuracy gains are $+1.22$ and $+5.42$ points for similar-event and same-topic cases; the previously noted $+4.78$ value corresponds to Macro-Precision, not Accuracy.

D.7 Hyperparameter Sensitivity

Table D.9 and Figure D.1 report the Strong/Weak threshold sweep, while Table D.10 and Figure 4 in the main paper reports routing-temperature and retrieval-size sensitivity.

As shown in Table D.9 and Figure D.1, the temporal-validation default is the best temporal configuration in the grid. The post-hoc event maximum at $(0.87, 0.70)$ routes 96.93% of samples to No, compared with 85.75% under the default, suggesting that unseen events may benefit from a more

Table D.9: Exact threshold-grid results on FakeSV (Macro-F1 mean ± standard deviation, %).

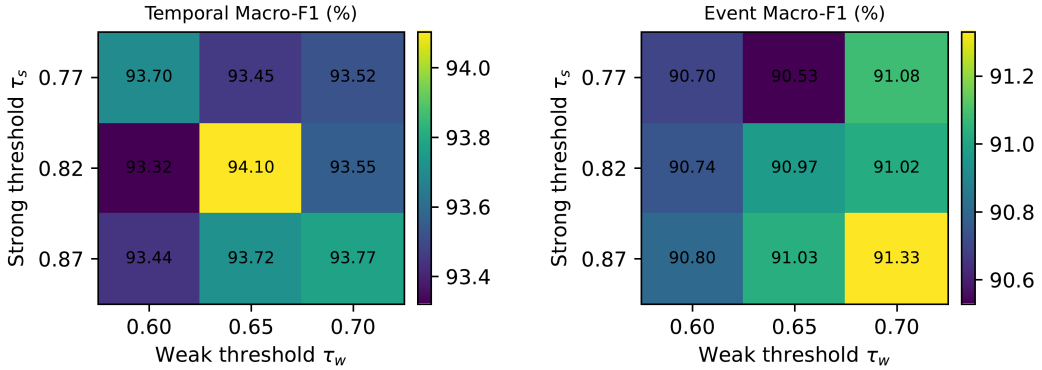
τ_s	τ_w	Temporal	Event
0.77	0.60	93.697 ± 0.215	90.696 ± 1.546
0.77	0.65	93.454 ± 0.458	90.528 ± 1.963
0.77	0.70	93.520 ± 0.816	91.080 ± 1.690
0.82	0.60	93.320 ± 0.354	90.742 ± 1.498
0.82	0.65	$\mathbf{94.103 \pm 0.513}$	90.968 ± 1.641
0.82	0.70	93.553 ± 0.652	91.023 ± 2.153
0.87	0.60	93.441 ± 0.445	90.802 ± 1.261
0.87	0.65	93.715 ± 0.351	91.030 ± 1.120
0.87	0.70	93.770 ± 0.842	$\mathbf{91.330 \pm 1.614}$

Table D.10: Routing-temperature and Top-k sensitivity on FakeSV (Macro-F1, %).

Configuration	Temporal	Event
$T_r = 0.03$	93.272	90.857
$T_r = 0.05$	93.608	91.050
$T_r = 0.10$ (default)	94.103	90.968
$T_r = 0.20$	93.919	90.668
$k = 1$	93.400	90.858
$k = 5$	93.546	90.803
$k = 10$ (default)	94.103	90.968
$k = 20$	93.659	91.304

Table D.11: Classification and calibration diagnostics on FakeSV. ECE, Brier score, and NLL are lower-is-better.

Split	Configuration	M-F1	ECE	Brier	NLL
Temporal	Full relation routing	94.103	0.034	0.052	0.241
Temporal	Random soft routing	93.774	0.053	0.057	0.351
Temporal	Random hard routing	93.445	0.050	0.059	0.298
Event	Full relation routing	90.968	0.051	0.074	0.311
Event	Random soft routing	90.637	0.053	0.078	0.383
Event	Random hard routing	91.168	0.067	0.079	0.408
Event	$(\tau_s, \tau_w) = (0.87, 0.70)$	91.330	0.060	0.078	0.408
Event	$k = 20$	91.304	0.054	0.074	0.340



(a) Temporal split

(b) Event split

Figure D.1: Strong/Weak threshold sensitivity on FakeSV. The default $(0.82, 0.65)$ is best temporally; several higher-threshold settings remain competitive on the event split within across-fold variation.

Table D.12: Paired PRME–TrueLens differences. Differences are percentage points; intervals are paired-bootstrap 95% CIs. Holm correction controls family-wise error across all 16 hypotheses, and BH controls false discovery rate.

Dataset	Split	Metric	Mean difference [95% CI]	Raw p	Holm p	BH q
FakeSV	Event	ACC	$+2.53$ $[+1.44, +3.63]$	.0152	.2122	.0498
FakeSV	Event	M-F1	$+2.54$ $[+1.43, +3.65]$	.0156	.2122	.0498
FakeSV	Event	M-P	$+2.48$ $[+1.53, +3.43]$	.0098	.1469	.0498
FakeSV	Event	M-R	$+2.53$ $[+1.44, +3.63]$	.0152	.2122	.0498
FakeSV	Temporal	ACC	$+0.74$ $[+0.33, +1.07]$	.0247	.2471	.0565
FakeSV	Temporal	M-F1	$+0.75$ $[+0.36, +1.07]$	.0202	.2224	.0539
FakeSV	Temporal	M-P	$+0.71$ $[+0.04, +1.23]$	.1010	.6058	.1469
FakeSV	Temporal	M-R	$+0.74$ $[+0.48, +0.98]$	.0070	.1123	.0498
FakeTT	Event	ACC	-0.25 $[-1.51, +1.01]$	.7454	1.0000	.8518
FakeTT	Event	M-F1	$+0.08$ $[-1.21, +1.29]$	.9180	1.0000	.9180
FakeTT	Event	M-P	-0.13 $[-1.23, +0.98]$	.8494	1.0000	.9061
FakeTT	Event	M-R	$+0.40$ $[-1.03, +1.48]$	.6094	1.0000	.7501
FakeTT	Temporal	ACC	$+2.07$ $[+0.74, +3.48]$	.0579	.5162	.1024
FakeTT	Temporal	M-F1	$+2.04$ $[+0.68, +3.51]$	.0640	.5162	.1024
FakeTT	Temporal	M-P	$+2.15$ $[+0.77, +3.61]$	.0574	.5162	.1024
FakeTT	Temporal	M-R	$+1.45$ $[+0.04, +2.76]$	.1330	.6652	.1774

conservative evidence-acceptance policy. Because this setting is identified on event test folds and its 0.361-point gain is below the fold standard deviation, it is not used to replace the unified default.

Table D.10 and Figure 4 in the main paper show that the model is not highly sensitive within the tested region. The default configuration is selected using the temporal validation protocol and is retained for all reported experiments; event-test means are not used for post-hoc hyperparameter selection. The event split sometimes favors higher thresholds or a larger candidate set, but these gains are smaller than the across-fold standard deviation.

D.8 Calibration and Routing Diagnostics

Table D.11 compares classification quality with ECE, Brier score, and NLL.

Random hard routing can match or slightly exceed the mean event-split M-F1 while producing substantially worse ECE and NLL. This result motivates reporting calibration together with classification performance: a route that occasionally chooses an incompatible evidence path may remain accurate on average but become overconfident when it fails.

D.9 Paired Statistical Results

Table D.12 reports the full family of 16 paired hypotheses: two datasets, two splits, and four aggregate metrics. Runs are paired by seed for temporal evaluation and by fold for event evaluation. Confidence intervals exhaustively enumerate all $5^5 = 3{,}125$ paired bootstrap resamples.

As shown in Table D.12, FakeSV event effects are consistently about 2.5 points and all paired-bootstrap intervals exclude zero; their BH-adjusted values are 0.0498. No comparison reaches 0.05 under the stricter Holm family-wise correction. FakeTT event effects are near zero with intervals spanning zero, whereas FakeTT temporal effect estimates are positive but uncertain. The exact two-sided sign-flip test has a minimum attainable $p = 0.0625$ with five pairs, so these results support consistent effect estimates but not a universal family-wise significance claim.

D.10 High-Similarity Non-Same-Event Subset

This diagnostic asks whether a highly similar precedent can still be misleading when it comes from a different event. The subset is generated only from temporal test queries. Among 264 queries whose Top-1 retrieval is not a same-event item, we select

$$\mathrm{PCS}_{max} \geq Q_{0.75}(\mathrm{PCS}_{max} \mid \text{non-same-event}), \tag{D.1}$$

$$\mathrm{TSS}_{max} \geq 0.50. \tag{D.2}$$

The resulting PCS cutoff is 0.706581 and yields 62 samples (27 fake, 35 real). The generation script first tries the 75th percentile, then 67th and 50th percentiles only when fewer than 60 samples remain,

Table D.13: Performance on the 62-sample high-similarity, non-same-event subset. Values are mean ± standard deviation across temporal seeds.

Method	ACC	M-F1	Fake recall	ECE	NLL
TrueLens (reproduced)	95.16 ± 1.14	95.03 ± 1.17	90.37	0.1082	0.2385
PRME	94.19 ± 0.88	94.05 ± 0.87	90.37	**0.0608**	**0.2370**

and caps the subset at 100 by descending PCS/TSS difficulty. This fallback was not activated for the reported subset.

As shown in Table D.13, PRME does not improve classification accuracy on this deliberately adversarial subset: its mean ACC and M-F1 are about one point lower, while fake recall is unchanged. Its ECE is substantially lower and NLL is similar. The supported conclusion is therefore narrower than a robustness claim: permission routing improves confidence calibration when highly similar but event-mismatched precedents are present, but the current subset does not demonstrate higher mean classification accuracy.

Table D.14: Complete comparison on FakeTT under temporal and event splits (%). † denotes unified baseline results reported by TrueLens; ∗ denotes our matched reproduction averaged over five temporal seeds and five event-isolated folds.

	Temporal				Event 5-fold			
Method	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
Unimodal								
BERT†	70.79	68.96	69.21	70.88	68.67	67.56	67.85	67.41
ViT†	64.32	64.15	65.58	66.41	70.24	68.15	69.54	68.61
Wav2Vec†	67.54	65.79	68.06	65.03	68.93	65.41	68.03	65.73
Conventional multimodal								
TikTec†	66.23	66.02	65.89	66.07	60.69	60.32	60.20	60.37
FANVM†	73.24	72.29	73.54	73.51	75.43	73.42	74.51	74.52
CAFE†	70.81	71.19	71.89	70.13	72.79	71.01	71.77	70.13
SV-FEND†	77.14	75.68	75.69	76.32	76.68	75.42	75.32	77.66
FakingRecipe†	80.01	78.23	78.19	79.31	75.25	73.90	75.34	74.24
ExMRD†	85.01	84.12	83.29	85.55	78.14	74.90	78.73	73.76
Retrieval-augmented								
REAL†	72.24	70.10	69.97	71.35	69.78	66.82	69.01	62.05
TrueLens∗ (reproduced)	85.55	84.41	83.52	86.24	78.12	76.86	77.29	76.76
PRME (Ours)	87.63	86.45	85.68	87.69	77.87	76.94	77.17	77.16

E FakeTT Data Analysis

E.1 Complete FakeTT Main Results

Table 1 gives the compact FakeTT ACC/Macro-F1 comparison, while Table D.14 restores Macro-Precision and Macro-Recall. PRME improves the reproduced TrueLens temporal Macro-F1 from 84.41 to 86.45, while the event-isolated means are effectively tied (76.86 versus 76.94).

E.2 Complete FakeTT Ablations

All differences in Tables E.1, E.2, and E.3 are computed directly from the displayed three-decimal full-model means: 86.450 for the temporal split and 76.939 for the event split. Positive event differences are bolded only to indicate a higher mean; they do not imply statistical significance because event-fold standard deviations are approximately 4–6 points.

Table E.1: FakeTT routing and modality ablations. Values are Macro-F1 mean ± standard deviation (%); parentheses give the difference from Full PRME.

Variant	Temporal	Event 5-fold
Routing controls		
Uniform branch weights	85.466 ± 0.697 (-0.984)	$\mathbf{77.540 \pm 5.817}$ $(\mathbf{+0.601})$
Rolled weights	83.865 ± 1.117 (-2.585)	$\mathbf{77.343 \pm 4.298}$ $(\mathbf{+0.404})$
Strong branch only	83.876 ± 1.029 (-2.574)	76.929 ± 5.508 (-0.010)
Weak branch only	83.804 ± 0.657 (-2.646)	$\mathbf{77.460 \pm 4.288}$ $(\mathbf{+0.521})$
No branch only	84.372 ± 1.065 (-2.078)	$\mathbf{77.632 \pm 5.238}$ $(\mathbf{+0.693})$
Hard relation routing	84.577 ± 1.545 (-1.873)	$\mathbf{77.092 \pm 4.419}$ $(\mathbf{+0.153})$
Random soft routing	84.166 ± 1.384 (-2.284)	$\mathbf{76.981 \pm 6.188}$ $(\mathbf{+0.042})$
Random hard routing	84.752 ± 1.895 (-1.698)	$\mathbf{77.348 \pm 5.805}$ $(\mathbf{+0.409})$
Modality controls		
Text only	84.764 ± 1.267 (-1.686)	76.066 ± 5.167 (-0.873)
Image only	76.971 ± 0.227 (-9.479)	73.199 ± 5.326 (-3.740)
Audio only	68.607 ± 1.153 (-17.843)	67.366 ± 5.195 (-9.573)
Text + image	84.233 ± 0.996 (-2.217)	$\mathbf{77.323 \pm 4.936}$ $(\mathbf{+0.384})$
Text + audio	84.879 ± 1.672 (-1.571)	76.905 ± 6.225 (-0.034)
Image + audio	77.098 ± 1.344 (-9.352)	72.506 ± 4.509 (-4.433)
Full PRME	$\mathbf{86.450 \pm 1.113}$	$\mathbf{76.939 \pm 5.218}$

Table E.2: FakeTT evidence, gating, fallback, and loss ablations (Macro-F1, %).

Variant	Temporal	Event 5-fold
Evidence and internal mechanism cues		
No Strong feature enhancement	84.653 ± 1.332 (-1.797)	$\mathbf{78.090 \pm 4.085}$ $(\mathbf{+1.151})$
No Strong external prior	84.933 ± 1.261 (-1.517)	76.757 ± 6.458 (-0.182)
No Weak external prior	84.818 ± 0.939 (-1.632)	76.695 ± 5.188 (-0.244)
No external priors	83.275 ± 1.014 (-3.175)	$\mathbf{77.125 \pm 4.800}$ $(\mathbf{+0.186})$
No Strong internal cue	85.157 ± 1.664 (-1.293)	$\mathbf{77.855 \pm 4.800}$ $(\mathbf{+0.916})$
No Weak internal cue	84.093 ± 1.077 (-2.357)	$\mathbf{77.615 \pm 4.606}$ $(\mathbf{+0.676})$
No No-route internal cue	84.159 ± 0.555 (-2.291)	$\mathbf{77.145 \pm 5.030}$ $(\mathbf{+0.206})$
No internal cues	84.749 ± 0.248 (-1.701)	76.754 ± 4.426 (-0.185)
Fallback, generic gating, and losses		
No fallback control	84.972 ± 1.936 (-1.478)	76.871 ± 6.204 (-0.068)
Mechanism-agnostic MoE	84.674 ± 1.257 (-1.776)	76.881 ± 5.996 (-0.058)
No mechanism supervision	85.020 ± 1.314 (-1.430)	$\mathbf{77.420 \pm 5.887}$ $(\mathbf{+0.481})$
No expert-balance loss	84.789 ± 1.353 (-1.661)	76.783 ± 5.480 (-0.156)
Full PRME	$\mathbf{86.450 \pm 1.113}$	$\mathbf{76.939 \pm 5.218}$

E.3 Split-Dependent Ablation Patterns

Across Tables E.1, E.2, and E.3, Full PRME reaches 86.450 ± 1.113 temporal Macro-F1 and almost every structural ablation decreases performance. Removing all external priors causes the largest gating-related decline (3.175 points), while Weak-only, mismatched routing, and Strong-only decrease by 2.646, 2.585, and 2.574 points. This pattern supports the value of sample-dependent routing, external evidence, and complementary branch fusion when temporally nearby precedents are available.

The same tables show that the event split behaves differently. Full PRME obtains 76.939 ± 5.218, while several simpler variants have higher means: the parameter-matched deep MLP reaches 78.543 ($+1.604$), removing Strong feature enhancement reaches 78.090 ($+1.151$), and removing the Strong internal cue reaches 77.855 ($+0.916$). All of these differences are much smaller than the 4–6 point across-fold standard deviations. The supported conclusion is therefore not that the simpler models are better, but that the complete architecture has not shown a stable event-isolated advantage on FakeTT and may be sensitive to event-specific external evidence.

E.4 Routing and Evidence Transfer

Table E.1 shows that random soft and random hard routing reduce temporal Macro-F1 by 2.284 and 1.698 points, and mismatched routing reduces it by 2.585 points. The temporal results therefore

Table E.3: FakeTT expert architecture, expert number, module, capacity, and hyperparameter controls (Macro-F1, %).

Variant	Temporal	Event 5-fold
Expert architecture		
One-layer MLP	84.505 ± 1.825 (-1.945)	$\mathbf{77.033 \pm 5.517}$ $(\mathbf{+0.094})$
Two-layer MLP	84.542 ± 1.132 (-1.908)	$\mathbf{77.354 \pm 5.028}$ $(\mathbf{+0.415})$
Three-layer MLP (default)	$\mathbf{86.450 \pm 1.113}$	$\mathbf{76.939 \pm 5.218}$
Four-layer MLP	84.447 ± 1.963 (-2.003)	76.868 ± 4.796 (-0.071)
Transformer expert	84.851 ± 1.650 (-1.599)	$\mathbf{77.333 \pm 5.445}$ $(\mathbf{+0.394})$
Expert number		
One expert (same architecture)	85.051 ± 0.920 (-1.399)	$\mathbf{77.183 \pm 5.012}$ $(\mathbf{+0.244})$
One expert (parameter matched)	84.988 ± 1.341 (-1.462)	76.685 ± 4.656 (-0.254)
Four merged-mechanism experts	85.331 ± 0.998 (-1.119)	76.872 ± 4.968 (-0.067)
Eight experts, no general expert	84.353 ± 0.887 (-2.097)	76.618 ± 4.602 (-0.321)
Nine experts (default)	$\mathbf{86.450 \pm 1.113}$	$\mathbf{76.939 \pm 5.218}$
Twelve experts	84.536 ± 1.261 (-1.914)	$\mathbf{76.978 \pm 4.946}$ $(\mathbf{+0.039})$
Module combinations		
Internal classifier only	85.083 ± 1.367 (-1.367)	$\mathbf{77.106 \pm 5.135}$ $(\mathbf{+0.167})$
Three paths + single head	84.284 ± 2.242 (-2.166)	76.487 ± 4.673 (-0.452)
Unified evidence + mechanism MoE	84.397 ± 0.823 (-2.053)	76.252 ± 5.835 (-0.687)
Three paths + mechanism MoE	$\mathbf{86.450 \pm 1.113}$	$\mathbf{76.939 \pm 5.218}$
Capacity controls		
Single internal branch (matched)	85.289 ± 0.950 (-1.161)	76.374 ± 5.045 (-0.565)
Deep MLP (matched)	83.923 ± 0.872 (-2.527)	$\mathbf{78.543 \pm 4.304}$ $(\mathbf{+1.604})$
Wide MLP (matched)	84.347 ± 1.896 (-2.103)	75.876 ± 4.575 (-1.063)
Hyperparameter sensitivity (Temporal only)		
Default/best configuration	$\mathbf{86.450 \pm 1.113}$	–
Routing temperature $T_r = 0.05$	84.475 ± 0.974 (-1.975)	–
Routing temperature $T_r = 0.20$	84.435 ± 1.411 (-2.015)	–
Retrieval size $k = 1$	84.403 ± 1.254 (-2.047)	–

support aligning route weights with the query's evidence condition. Under event isolation, however, random hard routing, uniform routing, and No-only have means 0.409, 0.601, and 0.693 points above Full PRME. Given the large fold variation, these configurations should be treated as comparable. The pattern suggests that precedents for unseen events are often unreliable and that a simpler or internal-evidence-dominant route may be more stable.

E.5 Modality Contribution

As shown in Table E.1, text is the strongest individual modality. Temporal Macro-F1 is 84.764 for text, 76.971 for image, and 68.607 for audio; audio-only is 17.843 points below the full model. Text+audio is the strongest two-modality temporal setting at 84.879 but still trails the full three-modality model. Under event isolation, text+image reaches 77.323 ($+0.384$), and text+audio is nearly tied with the full model. The robust interpretation is that text is the primary FakeTT signal, image provides useful complementary information, and audio is weak in isolation; the extra value of three-modality fusion is not stable across unseen events.

E.6 Expert Structure, Module Composition, and Capacity

Table E.3 shows that the default three-layer, nine-expert configuration is best among all tested expert depths and counts on the temporal split. Event-split means do not show a monotonic relationship with expert depth or expert number: two-layer MLP, Transformer experts, and the same-architecture single expert are all slightly above the full mean, but by substantially less than one fold standard deviation. Increasing expert complexity is therefore not uniformly beneficial under event shift.

In Table E.3, the temporal internal-only classifier reaches 85.083, so the complete model adds 1.367 points. In contrast, the three-path single-head and unified-evidence MoE variants are both below the internal-only classifier, indicating that the permission paths and mechanism experts are most effective when used jointly. On the event split, the internal-only model is slightly above Full PRME and the matched deep MLP has the highest mean in the table. These controls prevent the FakeTT temporal gain from being attributed solely to parameter count, while also exposing a limitation under strict event isolation.

E.7 HYPERPARAMETER SENSITIVITY

The hyperparameter rows in Table E.3 report only temporal FakeTT sweeps. The default configuration reaches 86.450 ± 1.113. Setting T_r to 0.05 or 0.20 reduces Macro-F1 by 1.975 and 2.015 points, and $k = 1$ reduces it by 2.047 points. The available evidence supports the default temporal configuration but does not establish that it is optimal under event isolation.

E.8 SUMMARY OF FAKETT FINDINGS

FakeTT event-isolated Macro-F1 is 9.511 points below the temporal result and has substantially larger variation. Permission-aware routing, external priors, and complete mechanism gating are consistently useful temporally, whereas the event split does not show a stable advantage over all simpler controls. The FakeTT result should therefore be presented as split dependent: PRME improves retrieval-augmented detection under temporal evaluation but remains statistically and practically comparable to TrueLens under event isolation.