ACHER: AVAILABILITY-CONDITIONED HIERARCHICAL EVIDENCE REASONING FOR MULTIMODAL FAKE NEWS DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fake news detection (MFND) identifies misinformation by jointly exploiting text, images, and their cross-modal relations, a task made increasingly hard as generative AI produces more complex cross-modal fakes. Existing methods extract veracity-discriminative signals via cross-modal alignment, consistency modeling, or feature fusion, yet typically assume sufficient corresponding evidence. In practice, low consistency either reflects conflict among existing evidence or insufficient evidence that leaves the relation unverifiable; confusing the two invites unfounded judgments. To address this, we propose the **A**vailability-**C**onditioned **H**ierarchical **E**vidence **R**easoner (ACHER), which conditions evidence reasoning on evidence availability: bidirectional sparse localization with null-hypothesis competition estimates it, relation reasoning proceeds only when evidence is available, and each direction yields compatible, conflicting, and unverified states. From these states, ACHER derives four complementary evidence mechanisms with state gating for joint veracity prediction. Experiments on Weibo, FineFake, and AMG show that ACHER consistently outperforms representative methods, confirming its effectiveness and generalization.

1 INTRODUCTION

Multimodal fake news detection has become crucial for identifying misinformation on social media, as generative AI increasingly produces diverse cross-modal fakes: paired text and images raise surface credibility while hiding deception behind factual distortion, image-text mismatch, content repurposing, or local manipulation. Real-world misinformation has likewise evolved from binary true/false claims toward complex manipulations and multi-type forgery (Chakraborty et al., 2023; Guo et al., 2025), so detection must go beyond intra-modal anomalies to judge whether text and image share reliable semantic correspondence.

Existing methods improve veracity discrimination by learning better joint representations or cross-modal interactions. For example, COOLANT aligns image-text relevance via contrastive learning (Wang et al., 2023); HSEN augments representations with text-relevant visual and knowledge semantics (Zhang et al., 2023); Event-Radar captures event-level text-image inconsistency (Ma et al., 2024); NLIN unifies images, text, and knowledge for natural-language reasoning (Zhang et al., 2024b); HAMI-M3D mines complementary clues from visual manipulation and latent intent (Wang et al., 2024); and BMLHF balances modalities via hierarchical fusion (Wu et al., 2025). Yet such models rarely ask whether the relation itself rests on sufficient corresponding evidence.

Low cross-modal consistency, however, carries no unique meaning. When the other modality contains corresponding content that clearly differs from the current one, low consistency signals conflict among existing evidence; when it lacks enough for reliable correspondence, it signals insufficient evidence. Prior studies noticed intra- and inter-modal uncertainty and content inconsistency (Wei et al., 2023; Sun et al., 2023), yet rarely told the two cases apart, conflating conflict with insufficiency and forcing relational judgments even when evidence is absent. Moreover, compatibility only means no obvious contradiction rather than factual support, and the two verification directions differ in candidate space, evidence granularity, and coverage, so their states should not be forced symmetric. Reasoning should therefore assess evidence availability first, then reason over the rela-

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.

tions of available evidence, and finally exploit the commonality and directionality of bidirectional states for veracity judgment.

To this end, we propose the Availability-Conditioned Hierarchical Evidence Reasoner (ACHER), which conditions cross-modal relation reasoning on evidence availability. Along the text-to-visual and visual-to-text directions, ACHER first localizes the most relevant corresponding evidence and contrasts its matching strength with an explicit null hypothesis, yielding a directional availability estimate. The resulting relation evidence is then interpreted through Availability-conditioned states, and each direction is organized into three mutually exclusive and closed states—compatible, conflicting, and unverified—where unverified reflects insufficient evidence rather than weak conflict. Exploiting the symmetric and antisymmetric structure of the bidirectional states, ACHER derives four complementary mechanisms (compatibility, conflict intensity, directional conflict, and unverified) and constrains the classifier to operate on state-gated mechanism representations rather than unconstrained global features.

In summary, our main contributions are:

1. We introduce an availability-conditioned perspective on cross-modal relations that separates conflict from insufficient evidence and distinguishes the conflicting and unverified states.
2. We propose ACHER, an availability-conditioned hierarchical evidence reasoning framework that estimates availability via bidirectional localization with null-hypothesis competition, forms compatible/conflicting/unverified states, and derives four complementary mechanisms for veracity prediction.
3. Experiments on Weibo, FineFake, and AMG show competitive or improved performance across heterogeneous multimodal fake-news benchmarks, while controlled interventions support the functional separation between evidence availability and conditional relation, and ablations establish the predictive utility of the structured components.

2 Related Works

2.1 Multimodal Fake News Detection

Multimodal fake news detection has evolved from feature fusion toward fine-grained cross-modal relation modeling. CAFE adaptively fuses unimodal and cross-modal cues via ambiguity estimation (Chen et al., 2022); BMR builds multi-perspective news representations from text, visual patterns, and semantics (Ying et al., 2023); CCD removes modality bias through causal intervention and counterfactual reasoning (Chen et al., 2023); and HAMMER jointly models veracity and local manipulation clues via shallow alignment and deep interaction (Shao et al., 2023). Knowledge- and LLM-augmented detectors, such as AKA-Fake (Zhang et al., 2024a), FKA-Owl (Liu et al., 2024), ERIC-FND (Cao et al., 2025), GLPN-LLM (Hu et al., 2025), and SDML (Zhu et al., 2025), further incorporate external knowledge or LLM assistance. These methods, however, generally do not explicitly model whether sufficient within-pair counterpart evidence is available before relation reasoning. ACHER instead estimates counterpart-evidence availability and conditions the resulting relation states on it.

2.2 Cross-modal Evidence and Relation Modeling

Beyond direct relation modeling, evidence retrieval and fact checking further strengthen verifiability: Abdelnabi et al. (2022) retrieve textual and visual evidence from web resources and verify claims through cross-modal cycle consistency; MOCHEG decomposes multimodal fact verification into evidence retrieval, claim verification, and explanation generation, with support, refute, or insufficient-evidence outcomes (Yao et al., 2023); Tahmasebi et al. (2024) exploit large language and vision-language models to retrieve and re-rank evidence; and SNIFFER detects out-of-context information by combining image-text consistency analysis with external retrieval (Qi et al., 2024), while COVE recovers the original context of an image to assist veracity judgment (Tonglet et al., 2025). In these studies, however, insufficient evidence means that retrieved information cannot support a factual verdict. Instead, ACHER conditions the compatible/conflicting distinction on whether

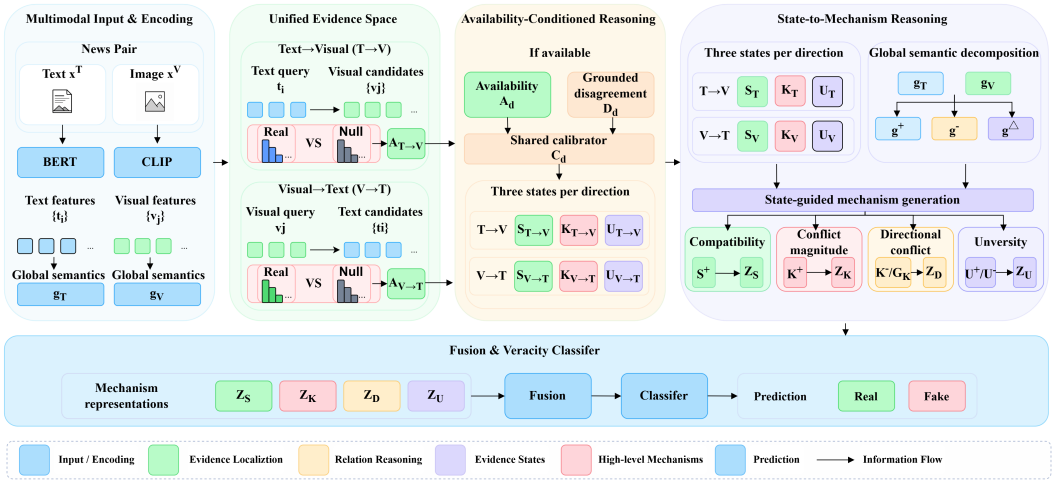


Figure 1: Overview of ACHER. Bidirectional sparse grounding with null-hypothesis competition first estimates counterpart-evidence availability. Availability-conditioned relation reasoning then identifies compatible, conflicting, and unverified states, which are organized by symmetric-antisymmetric factorization into four state-gated evidence mechanisms for veracity prediction.

corresponding evidence exists within the pair, with compatibility signaling no detected contradiction rather than factual support.

2.3 Hierarchical Multimodal Reasoning

Complex multimodal tasks increasingly adopt staged or structured reasoning rather than monolithic fusion. DDCoT splits visual recognition from language reasoning and integrates visual information over steps (Zheng et al., 2023); CCoT first generates a scene graph and reasons over it (Mitra et al., 2024); Visual CoT iteratively locates key visual regions to form intermediate reasoning (Shao et al., 2024); and Cantor adopts a perception-decision structure for visual acquisition and logical reasoning (Gao et al., 2024). This idea also appears in fake news detection, where T^{2}Agent organizes verification via tool calling and Monte Carlo tree search (Cui et al., 2026) and MMRGV mitigates the impact of erroneous reasoning via multi-perspective reasoning generation and cross-validation (Chen et al., 2026). Unlike hierarchies organized along perception duties, processing stages, or verification flows, ACHER's hierarchy is driven by the conditional dependency of evidence itself: availability is assessed first, relations are then inferred, and bidirectional states are organized into high-level mechanisms, each level feeding the next.

3 ACHER Framework

3.1 Overview

Given a multimodal news item with text $\boldsymbol{x}^T$ and image $\boldsymbol{x}^V$, we predict its veracity label y. Rather than classifying the joint representation directly, ACHER reasons over the evidence condition underlying the cross-modal relation, since low consistency may reflect conflict among existing corresponding evidence or merely a lack of sufficient corresponding content in the other modality. As shown in Fig. 1, ACHER separately estimates counterpart-evidence availability and relation discrepancy, then combines them into closed evidence states so that low Availability is assigned to Unverified rather than directly interpreted as conflict; the resulting states are further factorized into complementary evidence mechanisms with state-constrained information flow.

3.2 Unified Multimodal Evidence Space

ACHER adopts BERT and CLIP as text and visual encoders: token-level outputs and region features serve as local evidence, pooled representations supply global text/image semantics, and special text

tokens and the global visual token are excluded from local matching. With $\boldsymbol{t}_i, \boldsymbol{v}_j$ local and $\boldsymbol{t}_g, \boldsymbol{v}_g$ global, each modality is projected into a unified D-dimensional evidence space by an independent pre-normalized linear layer:

$$\boldsymbol{x}_i^T = \boldsymbol{P}_T(\boldsymbol{t}_i), \quad \boldsymbol{x}_j^V = \boldsymbol{P}_V(\boldsymbol{v}_j), \quad \boldsymbol{g}_T = \boldsymbol{P}_T(\boldsymbol{t}_g), \quad \boldsymbol{g}_V = \boldsymbol{P}_V(\boldsymbol{v}_g). \tag{1}$$

where $\boldsymbol{P}_T$ and $\boldsymbol{P}_V$ preserve modality-specific characteristics while the shared dimension keeps bidirectional localization and relation comparison at a consistent scale. Local representations establish cross-modal evidence relations; global representations act as evidence-state-constrained semantic carriers in the high-level mechanisms and are not directly used in local conflict judgment.

3.3 Bidirectional Sparse Evidence Localization

For either the text-to-visual or visual-to-text direction, indexed by d, let $\boldsymbol{x}_i^d$ be the i-th local source representation and $\boldsymbol{y}_j^d$ the j-th candidate in the other modality. Both are mapped into a matching space and sparse-normalized over real candidates only:

$$\boldsymbol{q}_i^d = \frac{\boldsymbol{W}_q \boldsymbol{x}_i^d}{\|\boldsymbol{W}_q \boldsymbol{x}_i^d\|}, \quad \boldsymbol{k}_j^d = \frac{\boldsymbol{W}_k \boldsymbol{y}_j^d}{\|\boldsymbol{W}_k \boldsymbol{y}_j^d\|}, \quad s_{ij}^d = \frac{(\boldsymbol{q}_i^d)^\top \boldsymbol{k}_j^d}{\tau_a}, \quad \alpha_i^d = \mathrm{sparsemax}\big(\boldsymbol{s}_i^d;\ \boldsymbol{m}^d\big). \tag{2}$$

where $\boldsymbol{m}^d$ masks valid candidates and τ_a is the matching temperature. This answers one conditional question: if counterpart evidence exists, where is it most likely located? The sparse weights aggregate the corresponding evidence at each source position, $\bar{\boldsymbol{y}}_i^d = \sum_j \alpha_{ij}^d \boldsymbol{y}_j^d$, whose matching energy is contrasted with an explicit null hypothesis. With $H(\alpha_i^d) \in [0, 1]$ the normalized entropy of the sparse distribution,

$$r_i^d = \sum_j \alpha_{ij}^d s_{ij}^d, \quad n_i^d = b_d + \gamma_d H(\alpha_i^d), \quad A_i^d = \sigma\big(r_i^d - n_i^d\big). \tag{3}$$

where r_i^d aggregates the real matching strength and n_i^d is a direction-specific null score with learnable scalars b_d and γ_d. Dispersion raises entropy and thus demands stronger real matching to beat the null hypothesis; with no valid candidate, A_i^d becomes zero. Hence A_i^d serves as the model's estimate of whether the other modality offers sufficient corresponding evidence for local judgment, rather than a direct text–image consistency score.

3.4 Availability-Conditioned Relation Decomposition

3.4.1 Evidence Discrepancy and Conditional Conflict

ACHER then judges the relation between each source content and its localized corresponding evidence. Because the two directions differ in candidate space and evidence granularity, each direction uses an independent relation descriptor: the source and localized representations are mapped and normalized to $\hat{\boldsymbol{x}}_i^d$ and $\hat{\boldsymbol{y}}_i^d$, capturing both the overall directional difference and per-dimension local deviations:

$$\begin{aligned}
\hat{\boldsymbol{x}}_i^d &= \frac{\boldsymbol{R}_d \boldsymbol{x}_i^d}{\|\boldsymbol{R}_d \boldsymbol{x}_i^d\|}, & \hat{\boldsymbol{y}}_i^d &= \frac{\boldsymbol{R}_d \bar{\boldsymbol{y}}_i^d}{\|\boldsymbol{R}_d \bar{\boldsymbol{y}}_i^d\|}, \\
D_{i,\cos}^d &= \tfrac{1}{2}\big(1 - (\hat{\boldsymbol{x}}_i^d)^\top \hat{\boldsymbol{y}}_i^d\big), & D_{i,\mathrm{abs}}^d &= \tfrac{1}{2} \sum_k w_k^d \big|\hat{\boldsymbol{x}}_{ik}^d - \hat{\boldsymbol{y}}_{ik}^d\big|, \\
D_i^d &= \beta_d D_{i,\cos}^d + (1 - \beta_d) D_{i,\mathrm{abs}}^d, & \sum_k w_k^d &= 1.
\end{aligned} \tag{4}$$

where $\boldsymbol{R}_d$, $\boldsymbol{w}^d$, and β_d are direction-specific parameters, and larger D_i^d indicates a more evident discrepancy between the source content and its localized evidence. A single scalar monotonic calibrator then maps both directions onto a common conditional-conflict scale, $c_i^d = \sigma[s(D_i^d - \theta)]$ with shared $s > 0$ and θ: direction-specific descriptors absorb cross-modal statistical differences, while the shared calibrator provides a comparable mapping across directions, with c_i^d interpreted as the model's conditional conflict estimate given sufficient corresponding evidence.

3.4.2 THREE-STATE EVIDENCE DECOMPOSITION

Availability A_i^d and conditional conflict probability c_i^d answer whether the relation can be verified and, if so, whether a conflict exists. They combine into three mutually exclusive and closed states:

$$S_i^d = A_i^d(1 - c_i^d), \quad K_i^d = A_i^d c_i^d, \quad U_i^d = 1 - A_i^d, \quad S_i^d + K_i^d + U_i^d = 1. \tag{5}$$

where S_i^d, K_i^d, and U_i^d are the compatible, conflicting, and unverified evidence qualities. Compatible means sufficient evidence exists without obvious conflict (non-contradiction, not factual support); conflicting means such evidence coexists with an obvious discrepancy between the source content and its localized evidence; unverified is defined directly by insufficient evidence, being neither weaker conflict nor low-confidence compatibility, so low A_i^d shifts quality to U_i^d.

3.4.3 STATE EVIDENCE AGGREGATION

State probabilities alone do not carry the content required by later reasoning. ACHER pools state-conditional local content with a generic state-weighted aggregation:

$$P(f \mid w, m) = \frac{\sum_i m_i w_i f_i}{\sum_i m_i w_i + \epsilon}. \tag{6}$$

where m is the valid-position mask; a state without valid evidence quality yields zero (its numerator also vanishes), avoiding tiny-denominator instability. Pooled content follows each state's semantics: compatible aggregates the shared content $(\boldsymbol{x}_i^d + \hat{\boldsymbol{y}}_i^d)/2$; conflicting keeps both the signed residual $\boldsymbol{x}_i^d - \hat{\boldsymbol{y}}_i^d$ and the direction-free difference $|\boldsymbol{x}_i^d - \hat{\boldsymbol{y}}_i^d|$; unverified aggregates only $\boldsymbol{x}_i^d$ itself, avoiding construction of a cross-modal difference when evidence is insufficient. Averaging state qualities over each direction's valid positions gives the direction-level evidence quality:

$$(S_d, K_d, U_d) = \frac{1}{N_d} \sum_i m_i^d (S_i^d, K_i^d, U_i^d), \quad N_d = \sum_i m_i^d. \tag{7}$$

ACHER thereby retains two groups of three-state evidence, one per direction, capturing whether each source modality obtains sufficient evidence from the other and how they relate; the two directions are not forced to share a state distribution, and their difference serves as structural information for later reasoning.

3.5 SYMMETRIC-ANTISYMMETRIC DECOMPOSITION

Simply concatenating the two directional states retains information but does not explicitly separate patterns shared by both directions from directional imbalance, so ACHER performs symmetric-antisymmetric decomposition at the evidence-state level. Let subscripts T and V denote the two directions:

$$S^+ = \frac{S_T + S_V}{2}, \quad K^\pm = \frac{K_T \pm K_V}{2}, \quad U^\pm = \frac{U_T \pm U_V}{2}. \tag{8}$$

where S^+ is the compatibility shared by both directions, K^+ the overall availability-conditioned conflict intensity, K^- the signed conflict difference, and U^+, U^- the overall unverified degree and its directional difference. To separate absolute directional difference from relative directional bias, we further define:

$$G_K = |K_T - K_V| = 2|K^-|, \quad B_K = \frac{K_T - K_V}{K_T + K_V + \epsilon}, \quad B_U = \frac{U_T - U_V}{U_T + U_V + \epsilon}. \tag{9}$$

where $G_K \in [0, 1]$ supplies the strength of the directional-conflict mechanism, and B_K, $B_U \in [-1, 1]$ indicate the dominating direction and the degree of relative imbalance.

3.6 STATE-CONSTRAINED HIGH-LEVEL MECHANISMS

To restrict the final classifier to evidence-state-conditioned information, ACHER uses the state qualities to gate information passing, with each mechanism scaled by its corresponding state strength. From the projected global text and visual representations we build their common, directional, and direction-free parts, $\boldsymbol{g}^+ = (\boldsymbol{g}_T + \boldsymbol{g}_V)/2$, $\boldsymbol{g}^- = (\boldsymbol{g}_T - \boldsymbol{g}_V)/2$, and $\boldsymbol{g}^\Delta = |\boldsymbol{g}_T - \boldsymbol{g}_V|$. Let $\boldsymbol{e}_d^S$,

$\boldsymbol{e}_d^{K,m}$, $\boldsymbol{e}_d^{K,s}$, and $\boldsymbol{e}_d^{U}$ denote the direction-level compatible content, direction-free conflict magnitude, signed conflict residual, and unverified content of the previous section. ACHER applies the direction-level state constraints before combining the two directions:

$$\begin{aligned} l^S = \tfrac{1}{2}\big(S_T \boldsymbol{e}_T^S + S_V \boldsymbol{e}_V^S\big), \qquad l^K = \tfrac{1}{2}\big(K_T \boldsymbol{e}_T^{K,m} + K_V \boldsymbol{e}_V^{K,m}\big), \\ l^D = \tfrac{1}{2}\big(K_T \boldsymbol{e}_T^{K,s} - K_V \boldsymbol{e}_V^{K,s}\big), \\ l^{U,+} = \tfrac{1}{2}\big(U_T \boldsymbol{e}_T^{U} + U_V \boldsymbol{e}_V^{U}\big), \qquad l^{U,-} = \tfrac{1}{2}\big(U_T \boldsymbol{e}_T^{U} - U_V \boldsymbol{e}_V^{U}\big). \end{aligned} \tag{10}$$

Global semantic carriers follow the same principle:

$$\begin{aligned} \tilde{\boldsymbol{g}}^S = S^+ \boldsymbol{g}^+, \quad \tilde{\boldsymbol{g}}^K = K^+ \boldsymbol{g}^\Delta, \quad \tilde{\boldsymbol{g}}^D = G_K \boldsymbol{g}^-, \\ \tilde{\boldsymbol{g}}^{U,+} = \tfrac{1}{2}\big(U_T \boldsymbol{g}_T + U_V \boldsymbol{g}_V\big), \quad \tilde{\boldsymbol{g}}^{U,-} = \tfrac{1}{2}\big(U_T \boldsymbol{g}_T - U_V \boldsymbol{g}_V\big). \end{aligned} \tag{11}$$

As with local content, the unverified global carrier does not rebuild cross-modal relations; it keeps each direction's own semantics and directional differences when corresponding evidence is insufficient. The four mechanisms are realized by independent lightweight nonlinear mappings, and the per-mechanism strength $g_m \in \{S^+, K^+, G_K, U^+\}$ comes from equation 8–equation 9:

$$\tilde{\boldsymbol{z}}^m = \Phi_m\big(\,[\,l^m;\ \tilde{\boldsymbol{g}}^m;\ \eta^m\,]\,\big), \quad \boldsymbol{z}^m = g_m \tilde{\boldsymbol{z}}^m. \tag{12}$$

For the unverified mechanism, l^U and $\tilde{\boldsymbol{g}}^U$ carry the common and directional terms of equation 10 and equation 11, and η^U further keeps U_T, U_V, U^-, B_U, and the localization uncertainty. ACHER thus adopts two-level information control: each direction's local content and global carriers are state-constrained before entering a mechanism, and gating is applied again by mechanism strength after the nonlinear mapping. The classifier operates on the state-constrained representations $\boldsymbol{z}^S, \boldsymbol{z}^K, \boldsymbol{z}^D, \boldsymbol{z}^U$ without directly using the unconstrained global features $\boldsymbol{g}_T$ or $\boldsymbol{g}_V$, so state quality serves as a structural constraint on information flow rather than merely an additional statistical feature.

3.7 Training Objective

We train the mechanisms with label-smoothed cross-entropy $\mathcal{L}_{\text{cls}}$ and two auxiliary objectives that encourage mechanism-specific structure. A state-conditioned auxiliary loss supervises each mechanism with a lightweight head $C_m(\cdot)$ in proportion to its evidence-state quality:

$$\mathcal{L}_{\text{aux}} = \mathbb{E}_{b,m}\big[\text{sg}(g_{m,b})\,\ell_{\text{cls}}\big(C_m(\boldsymbol{z}_b^m),\, y_b\big)\big]\,. \tag{13}$$

where $\mathbb{E}_{b,m}$ averages over the batch and mechanisms, g_m is the mechanism strength (equation 12), and $\text{sg}(\cdot)$ stops gradients. To discourage a mechanism from behaving as an unconstrained feature branch, a structural semantic loss uses a training-only branch that reuses Φ_m without the state-magnitude gating or state statistics, $\hat{\boldsymbol{z}}^m = \Phi_m([\hat{l}^m; \hat{\boldsymbol{g}}^m; 0])$, and is trained to recover the structural targets $t^S{=}S^+$, $t^K{=}K^+$, $t^D{=}[B_K, G_K]$, $t^U{=}[U^+, B_U]$:

$$\mathcal{L}_{\text{sem}} = \frac{1}{6}\sum_{r\in R} \mathbb{E}_b\big[\text{sg}(\omega_r)\,\ell\big(\hat{t}_r,\, \text{sg}(t_r)\big)\big]\,. \tag{14}$$

where $R = \{S^+, K^+, B_K, G_K, U^+, B_U\}$; $\omega_{B_K} = K^+$ and $\omega_{B_U} = U^+$ weight the direction-balance targets, and $\omega_r = 1$ otherwise. Without state scalars or gating-induced magnitudes to exploit, recovering these targets encourages the mechanism content to remain aligned with the corresponding structural quantities. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{sem}}\rho(e)\,\mathcal{L}_{\text{sem}}, \quad \rho(e) = \min(1,\, e/3). \tag{15}$$

where e counts training epochs, first stabilizing veracity discrimination before strengthening the mechanism–evidence correspondence; the auxiliary and structural heads are training-only and dropped at inference, so prediction is based on the four state-constrained representations $\boldsymbol{z}^m$.

4 Experiments

4.1 Datasets and Implementation

We evaluate ACHER on three multimodal fake-news benchmarks. Weibo contains 9,528 image–text items (4,779 real and 4,749 fake) and follows the official train/validation/test IDs (Jin et al.,

Table 1: Performance comparison on Weibo, FineFake, and AMG. Best and second-best results are shown in **bold** and underlined.

Method	Weibo			FineFake			AMG	
	ACC ↑	$F1_F$ ↑	$F1_R$ ↑	ACC ↑	$F1_F$ ↑	$F1_R$ ↑	ACC ↑	M-F1 ↑
SpotFake	0.892	0.932	0.739	0.775	0.762	0.779	–	–
CAFE	0.840	0.842	0.837	0.787	0.769	0.774	0.7667	0.7628
CMC	0.893	0.899	0.907	0.790	0.775	0.786	–	–
BMR	0.918	0.914	0.904	0.792	0.772	0.789	0.8079	0.8057
MRHFR	0.922	0.927	0.910	0.781	0.765	0.782	–	–
DAPT	0.944	0.948	0.939	**0.809**	0.785	**0.828**	–	–
CLIP	–	–	–	–	–	–	0.7812	0.7809
MCAN	–	–	–	–	–	–	0.7740	0.7693
MGCA	–	–	–	–	–	–	0.8323	0.8310
ACHER (ours)	**0.949**	**0.951**	**0.947**	**0.809**	**0.829**	0.782	**0.8550**	**0.8509**

2017). FineFake contains 16,909 items (10,507 real and 6,402 fake) under a 6:2:2 split (Zhou et al., 2026). AMG contains 5,022 items (3,018 real and 2,004 fake), covering image forgery, Ineffective Visual Information, and entity-, event-, and temporal-level inconsistencies (Guo et al., 2025). Weibo and FineFake are used for comparison, ablation, and mechanistic analysis; AMG serves as a complementary comparison benchmark.

ACHER uses BERT text features and CLIP visual features and is trained in PyTorch with batch size 32 for at most 50 epochs using AdamW, one warmup epoch, cosine learning-rate decay, and validation-based early stopping. The objective uses label smoothing 0.02 and auxiliary/structural weights 0.10/0.08. Additional architecture and training details are given in Appendix A.1.

All reported experiments use random seed 42; to assess training stability, we additionally repeat the full ACHER training and evaluation with seeds 666 and 999, and report the resulting three-seed statistics in Appendix A.2.

4.2 Comparison Study

We compare ACHER with representative multimodal fake-news detectors on Weibo, FineFake, and AMG. The primary-dataset baselines include SpotFake (Singhal et al., 2019), CAFE (Chen et al., 2022), CMC (Wei et al., 2022), BMR (Ying et al., 2023), MRHFR (Wu et al., 2023), and DAPT (Tong et al., 2025); on AMG we additionally include CLIP (Radford et al., 2021), MCAN (Wu et al., 2021), and MGCA (Guo et al., 2025).

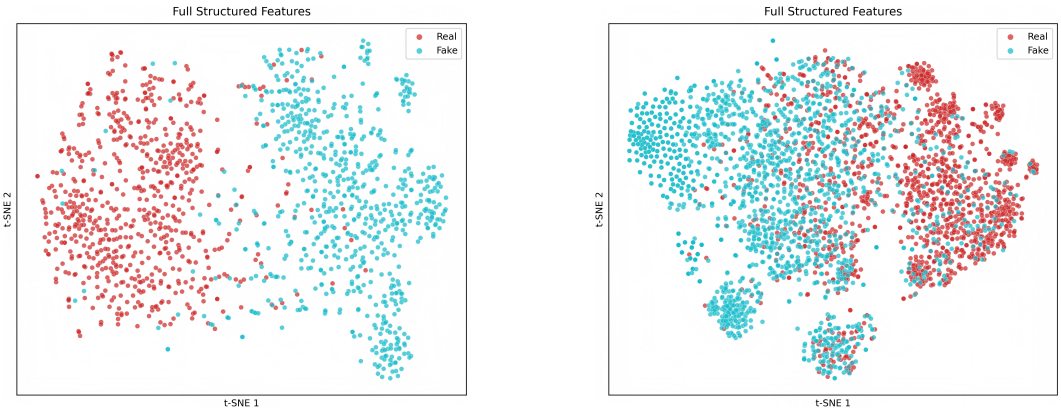
On Weibo, ACHER obtains 0.949 ACC and 0.951/0.947 fake/real F1, improving over DAPT by 0.5/0.3/0.8 percentage points. On FineFake, it matches DAPT at 0.809 ACC with 0.829/0.782 fake/real F1. We additionally examine class balance and operating-point sensitivity on both Weibo and FineFake. At the default threshold, the predicted fake proportions closely track the empirical class rates on both datasets; on FineFake, the validation-selected balanced-accuracy threshold is 0.493, close to the default 0.5. Full decision-boundary and ranking diagnostics are reported in Appendix A.3. On AMG, ACHER reaches 0.8550 ACC and 0.8509 Macro-F1, exceeding MGCA by 2.27/1.99 points. Overall, ACHER is competitive or improved across the three heterogeneous benchmarks.

4.3 Ablation Study

Table 2 retains the ablations most directly tied to ACHER's two-level decomposition; the complete matrix is in Appendix A.4. Availability, null competition, sparse localization, and both directions all improve performance. Replacing symmetric–antisymmetric decomposition with direct directional-state concatenation causes the largest drop (0.949→0.931 on Weibo; 0.809→0.785 on FineFake), while removing any mechanism or evidence-state gating also degrades both datasets.

Table 2: Core ablations (ACC). Full ablations are in Appendix A.4.

Method	Weibo	FineFake	Method	Weibo	FineFake
ACHER (full)	**0.949**	**0.809**			
W/o evidence availability	0.942	0.804	W/o sym.–antisym. decomp.	0.931	0.785
W/o null competition	0.943	0.805	W/o compatibility mech.	0.942	0.794
W/o sparse localization	0.941	0.801	W/o conflict-mag. mech.	0.938	0.801
Text-to-visual only	0.939	0.798	W/o directional-conflict	0.945	0.802
Visual-to-text only	0.941	0.804	W/o unverified mech.	0.945	0.804
			W/o evidence-state gating	0.942	0.799



(a) Weibo (b) FineFake

Figure 2: t-SNE visualization of ACHER's full structured representations on Weibo and FineFake.

Representation visualization. Figure 2 provides a qualitative view of the full structured representation on both datasets; the controlled interventions below further evaluate the functional roles of Availability and conditional Relation. Mechanism-specific views are in Appendix A.5.

4.4 Intervention-Based Mechanistic Analysis

Ablations quantify the predictive contribution of the structured components. We further probe *intervention selectivity* with frozen checkpoints on the Weibo and FineFake validation splits: evidence removal should reduce Availability, whereas counterpart replacement should alter Relation even when Availability is controlled. All diagnostic samples contain a physically present image and non-empty processed text; the exact audit counts are reported in Appendix A.6.

Targeted evidence deletion. For each direction, we delete the top $r \in \{5, 10, 15, 20\}\%$ counterpart candidates ranked by received sparse-grounding mass and compare against matched random deletion. Table 3 reports the full dose sequence: targeted deletion monotonically lowers A in every dataset–direction pair and is substantially stronger than random deletion. At 20%, the targeted effect is 27.8×/7.3× stronger on Weibo and 15.8×/5.4× stronger on FineFake for $T \to V$/$V \to T$, respectively. Full U responses, state-conservation checks, and the prediction-direction analysis are given in Appendix A.7.

Availability-controlled counterpart replacement. We next keep the source fixed and replace the counterpart with a content-distinct sample, averaging four content-aware derangements. The main text reports the strict cohort, $|A^{\mathrm{cf}} - A^{\mathrm{obs}}| < 0.01$, while the full sensitivity analysis is provided in Appendix A.8. As Table 4 shows, Mean $|\Delta A|$ is at most 0.0054, yet D and c increase in all four dataset–direction pairs; the induced S decrease and K increase are consistent with the closed-state construction.

Table 3: Evidence deletion. Each sequence reports $100\Delta A$ at $r = 5/10/15/20\%$; Tar./Rand. denote targeted/matched-random deletion.

Dataset	Direction	$100\Delta A_{\text{Tar.}}$	$100\Delta A_{\text{Rand.}}$
Weibo	$T \to V$	$-2.98/-4.65/-5.50/-6.16$	$-0.05/-0.10/-0.15/-0.22$
Weibo	$V \to T$	$-1.36/-2.06/-2.60/-3.06$	$-0.10/-0.20/-0.30/-0.42$
FineFake	$T \to V$	$-1.72/-2.52/-2.95/-3.34$	$-0.05/-0.10/-0.15/-0.21$
FineFake	$V \to T$	$-1.02/-1.50/-1.94/-2.37$	$-0.13/-0.21/-0.32/-0.44$

Table 4: Strict Availability-controlled counterpart replacement ($|A^{\text{cf}} - A^{\text{obs}}| < 0.01$). Δ denotes counterfactual minus observed.

Dataset	Direction	N	Mean $\lvert\Delta A\rvert$	ΔD	Δc	ΔS	ΔK
Weibo	$T \to V$	205	0.0039	$+0.0047$	$+0.0068$	-0.0046	$+0.0048$
Weibo	$V \to T$	67	0.0054	$+0.0075$	$+0.0086$	-0.0071	$+0.0046$
FineFake	$T \to V$	402	0.0046	$+0.0035$	$+0.0049$	-0.0044	$+0.0028$
FineFake	$V \to T$	67	0.0045	$+0.0128$	$+0.0180$	-0.0127	$+0.0114$

Mechanistic summary and robustness. The two interventions target different levels of the hierarchy: deletion changes whether localized counterpart evidence remains sufficient, whereas replacement changes how an available counterpart relates to the fixed source while approximately preserving Availability. This supports the intended functional separation between Availability and conditional Relation. An independent strict-matching rerun with 2,000 paired bootstrap resamples yields consistent effects across all four dataset–direction pairs; full confidence intervals and tolerance sensitivity are in Appendix A.9.

Evidence-regime analysis. We additionally ask whether the trained classifier uses Unverified preferentially when model-estimated counterpart evidence is scarce. Validation medians define frozen both-high, asymmetric, and both-low Availability cohorts; within each cohort we zero only $\boldsymbol{z}^U$. Table 5 reports the resulting Macro-F1 change, while full cohort statistics and all mechanism knockouts are in Appendix A.10.

Table 5: Unverified dependence across Availability regimes. Entries are $\Delta_{\neg U}$ in Macro-F1 percentage points (knockout minus full); more negative values indicate stronger dependence on Unverified.

Dataset	Both-high	Asymmetric	Both-low
FineFake	$+0.06$	-0.11	-0.52
Weibo	-0.45	-0.92	-2.35

On both datasets, the Unverified knockout becomes progressively more harmful from both-high to asymmetric to both-low Availability. Thus, within ACHER's validation-defined partition, reliance on Unverified increases as model-estimated counterpart-evidence availability decreases, linking evidence sufficiency to state-dependent classifier use.

5 Conclusion

In this paper, we introduced ACHER for multimodal fake news detection under ambiguous cross-modal evidence. ACHER explicitly models counterpart-evidence availability before relation judgment, distinguishing compatible, conflicting, and unverified states and organizing bidirectional relations into complementary evidence mechanisms. Experiments on Weibo, FineFake, and AMG show competitive or improved performance across heterogeneous multimodal fake-news benchmarks, while controlled ablations and interventions support the intended separation between evidence availability and conditional relation. These results suggest that separating evidence sufficiency from relation judgment is a useful direction for structured cross-modal veracity reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance (text polishing and grammar checking of selected paragraphs). We have not used generative AI tools for ideation, experimental design, implementation, or data analysis. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

No human subjects were involved in this study. All datasets used (Weibo, FineFake, AMG) are publicly available benchmarks and were used in accordance with their licenses and intended purpose. The method targets misinformation detection, which is intended to help mitigate social harm from fake news. We did not identify additional dataset- or method-specific ethical concerns beyond those associated with misinformation detection and benchmark use.

REPRODUCIBILITY STATEMENT

We describe the architecture, training schedule, hyperparameters, and data splits in detail in Sections 3 and 4. Pretrained encoders (BERT, CLIP) and training framework (PyTorch) are publicly available. We will release the source code and trained checkpoints upon acceptance. Datasets (Weibo, FineFake, AMG) are publicly available.

REFERENCES

S. Abdelnabi, R. Hasan, and M. Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14940–14949, 2022.

B. Cao, Q. Wu, J. Cao, B. Liu, and J. Gui. External reliable information-enhanced multimodal contrastive learning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 31–39, 2025.

M. Chakraborty, K. Pahwa, A. Rani, S. Chatterjee, D. Dalal, H. Dave, R. G, P. Gurumurthy, A. Mahor, S. Mukherjee, A. Pakala, I. Paul, J. Reddy, A. Sarkar, K. Sensharma, A. Chadha, A. Sheth, and A. Das. FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W question-answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15282–15322, 2023.

J. Chen, Y. Li, K. C. Ng, H. Wang, and L.-J. Zhang. Toward multimodal fake news detection by multi-perspective rationale generation and verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 66–74, 2026.

Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, T. Lu, and L. Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, pp. 2897–2905, 2022.

Z. Chen, L. Hu, W. Li, Y. Shao, and L. Nie. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 627–638, 2023.

X. Cui, Y. Zou, Z. Li, P. Li, X. Xu, X. Liu, and H. Huang. T2Agent: A tool-augmented multimodal misinformation detection agent with Monte Carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 175–183, 2026.

T. Gao, P. Chen, M. Zhang, C. Fu, Y. Shen, Y. Zhang, S. Zhang, X. Zheng, X. Sun, L. Cao, and R. Ji. Cantor: Inspiring multimodal chain-of-thought of MLLM. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9096–9105, 2024.

H. Guo, Z. Ma, Z. Zeng, M. Luo, W. Zeng, J. Tang, and X. Zhao. Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 228–236, 2025.

S. Hu, J. Hu, and H. Zhang. Synergizing LLMs with global label propagation for multimodal fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1440, 2025.

Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816, 2017.

X. Liu, P. Li, H. Huang, Z. Li, X. Cui, J. Liang, L. Qin, W. Deng, and Z. He. FKA-Owl: Advancing multimodal fake news detection through knowledge-augmented LVLMs. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10154–10163, 2024.

Z. Ma, M. Luo, H. Guo, Z. Zeng, Y. Hao, and X. Zhao. Event-Radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5809–5821, 2024.

C. Mitra, B. Huang, T. Darrell, and R. Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024.

P. Qi, Z. Yan, W. Hsu, and M. L. Lee. SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13052–13062, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems 37*, pp. 8612–8642, 2024.

R. Shao, T. Wu, and Z. Liu. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2023.

S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data*, pp. 39–47, 2019.

M. Sun, X. Zhang, J. Ma, S. Xie, Y. Liu, and P. S. Yu. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12736–12749, 2023.

S. Tahmasebi, E. Müller-Budack, and R. Ewerth. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2189–2199, 2024.

Y. Tong, W. Lu, X. Cui, Y. Mao, and Z. Zhao. DAPT: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7902–7911, 2025.

J. Tonglet, G. Thiem, and I. Gurevych. COVE: COntext and VEracity prediction for out-of-context images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2029–2049, 2025.

B. Wang, S. Wang, C. Li, R. Guan, and X. Li. Harmfully manipulated images matter in multimodal misinformation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2262–2271, 2024.

L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, and S. Wang. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5696–5704, 2023.

L. Wei, D. Hu, W. Zhou, and S. Hu. Modeling both intra- and inter-modality uncertainty for multimodal fake news detection. *IEEE Transactions on Multimedia*, 25:7906–7916, 2023.

Z. Wei, H. Pan, L. Qiao, X. Niu, P. Dong, and D. Li. Cross-modal knowledge distillation in multimodal fake news detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4733–4737, 2022.

F. Wu, S. Chen, G. Gao, Y. Ji, and X.-Y. Jing. Balanced multi-modal learning with hierarchical fusion for fake news detection. *Pattern Recognition*, 164:111485, 2025.

L. Wu, P. Liu, and Y. Zhang. See how you read? multi-reading habits fusion reasoning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13736–13744, 2023.

Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2560–2569, 2021.

B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2733–2743, 2023.

Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5384–5392, 2023.

L. Zhang, X. Zhang, Z. Zhou, F. Huang, and C. Li. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16777–16785, 2024a.

Q. Zhang, J. Liu, F. Zhang, J. Xie, and Z.-J. Zha. Hierarchical semantic enhancement network for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3424–3433, 2023.

Q. Zhang, J. Liu, F. Zhang, J. Xie, and Z.-J. Zha. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 2542–2550, 2024b.

G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems 36*, pp. 5168–5191, 2023.

Z. Zhou, X. Zhang, L. Zhang, J. Liu, E. Cambria, and C. Li. FineFake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection. *Information Fusion*, 132:104253, 2026.

Y. Zhu, Y. Wang, and Z. Yu. Multimodal fake news detection: MFND dataset and shallow-deep multitask learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8012–8020, 2025.

A APPENDIX

This appendix provides additional implementation details, diagnostic analyses, robustness checks, and representation visualizations for ACHER. The appendix is organized as follows:

- §A.1 Implementation Details.
- §A.2 Multi-Seed Stability Evaluation.
- §A.3 Decision-Boundary and Class-Balance Audit.
- §A.4 Complete Ablation Results.
- §A.5 Additional Representation Visualizations.
- §A.6 Physical-Presence Audit for Mechanistic Diagnostics.
- §A.7 Full Evidence-Deletion Results.
- §A.8 Availability-Matching Sensitivity.
- §A.9 Independent Bootstrap Robustness Details.
- §A.10 Evidence-Regime Analysis Details.

A.1 IMPLEMENTATION DETAILS

The shared evidence-reasoning dimension is 256, the cross-modal attention key dimension is 128 with temperature 0.45, each of the four mechanism channels has dimension 128, and the interaction classifier uses hidden dimensions $[256, 128, 64]$. We use CLIP ViT-B/16 as the visual encoder for all three datasets. For text encoding, Weibo uses `bert-base-chinese`, while FineFake and AMG use `bert-base-uncased`. We fine-tune the last three encoder layers of both the text and visual encoders while keeping the remaining pretrained layers frozen. The maximum text lengths are 160, 197, and 192 tokens for Weibo, FineFake, and AMG, respectively, and images are processed at 224×224 resolution.

Training uses AdamW with parameter-group learning rates of 8×10^{-6} for the text encoder, 1.5×10^{-6} for the image encoder, 2.5×10^{-5} for the modality projections, 1.5×10^{-5} for cross-modal localization, 2.25×10^{-5} for the relation descriptors and conflict calibrator, 3×10^{-5} for the mechanism experts, 2×10^{-5} for the structural-semantic heads, and 5×10^{-5} for the final classifier. Weight decay is 0.01 for the text encoder, modality projections, cross-modal localization, mechanism experts, and structural-semantic heads; 0.02 for the image encoder and classifier; and 0.025 for the relation descriptors, while the conflict calibrator and bias/one-dimensional parameters use no weight decay. We use batch size 32, one warmup epoch followed by cosine learning-rate decay, and train for at most 50 epochs. Checkpoints are selected by validation Macro-F1 with early-stopping patience 6.

The classification objective uses label smoothing 0.02, and the mechanism-level auxiliary and structural-semantic losses are weighted 0.10 and 0.08, respectively, with the structural term ramped in over the first three epochs. Training uses automatic mixed precision, with gradient norms clipped to 0.35 for the cross-modal localization module and 1.0 globally. We use random seed 42 with deterministic CUDA operations enabled. Experiments are conducted on a single NVIDIA GeForce RTX 4060 Ti GPU with 16 GB memory using Python 3.10.16, PyTorch 2.6.0 with CUDA 11.8, Transformers 4.48.3, and TorchVision 0.21.0.

A.2 MULTI-SEED STABILITY EVALUATION

All results reported in the main paper are obtained with random seed 42. To evaluate sensitivity to training stochasticity, we additionally rerun the full ACHER training and evaluation pipeline with random seeds 666 and 999 on Weibo, FineFake, and AMG, yielding three runs per dataset with seeds $\{42, 666, 999\}$. Across seeds, we keep the dataset splits, preprocessing, pretrained encoders, model architecture, optimization hyperparameters, learning-rate schedules, early-stopping rule, checkpoint-selection criterion, and test-time evaluation protocol unchanged. Deterministic CUDA operations remain enabled in every run, and only the random seed is changed.

Table 6: Multi-seed stability of ACHER on Weibo, FineFake, and AMG. Mean and standard deviation are computed over seeds 42, 666, and 999.

Seed	Weibo			FineFake			AMG	
	ACC↑	$F1_F$ ↑	$F1_R$ ↑	ACC↑	$F1_F$ ↑	$F1_R$ ↑	ACC↑	M-F1↑
42	0.949	0.951	0.947	0.809	0.829	0.782	0.8550	0.8509
666	0.944	0.945	0.944	0.808	0.841	0.761	0.8471	0.8441
999	0.947	0.947	0.946	0.804	0.831	0.765	0.8537	0.8511
Mean	0.947	0.948	0.946	0.807	0.834	0.769	0.8519	0.8487
Std.	0.003	0.003	0.002	0.003	0.006	0.011	0.0042	0.0040

Table 7: Decision-boundary and class-balance audit at the default threshold. Pred. Fake and True Fake denote the predicted and empirical fake-class proportions. AUROC and AP are threshold-independent ranking metrics.

Dataset	Macro-F1	Bal. Acc.	Pred. Fake	True Fake	AUROC	AP_F	AP_R
Weibo	0.9493	0.9491	0.5212	0.5144	0.9819	0.9756	0.9842
FineFake	0.8057	0.8061	0.5576	0.5635	0.8861	0.9139	0.8503

For each run, checkpoint selection is performed independently using validation Macro-F1 under the same early-stopping protocol described in Appendix A.1; the selected checkpoint is then evaluated on the corresponding test split. The seed-42 runs remain the source of the single-run numbers reported in the main paper, while the additional seeds are used solely to quantify stability rather than to select a favorable result. Table 6 reports the per-seed metrics together with the mean and standard deviation across the three runs.

A.3 Decision-Boundary and Class-Balance Audit

The main comparison uses the default decision threshold of 0.5. We further analyze class balance and operating-point sensitivity on both Weibo and FineFake. Table 7 reports default-threshold diagnostics for both datasets. Because FineFake exhibits greater class-wise F1 asymmetry, we additionally select a balanced-accuracy threshold on its validation split and freeze it before test evaluation, with no threshold tuning on the test set.

Class-rate and validation-threshold checks. At the default operating point, the predicted fake proportion closely tracks the empirical class prevalence: the gap is $+0.68$ percentage points on Weibo and -0.59 points on FineFake. On FineFake, the validation audit evaluates 1,001 equally spaced thresholds in $[0, 1]$ and selects the threshold that maximizes balanced accuracy; ties are resolved by smaller absolute fake-rate gap, then by proximity to 0.5, and finally by the smaller threshold. This procedure gives 0.493, which is close to the default 0.5. Freezing this threshold and applying it to the test set yields 0.8072 Macro-F1 and predicts 56.29% fake samples versus a true rate of 56.35%. The validation-selected threshold therefore remains close to 0.5 and produces similar class proportions and Macro-F1 on the test set.

Threshold-independent ranking and logit-bias control. A class-specific logit offset can alter the operating point, and hence class-wise performance and the predicted class rate, without improving the underlying ranking. We therefore complement the default-threshold audit with the threshold-independent AUROC and class-wise AP metrics in Table 7, and further perform a frozen-checkpoint logit-bias control. Specifically, we add a constant bias δ to the fake-class logit while keeping all learned representations and model parameters fixed. The resulting test-set performance is reported in Table 8.

Increasing the fake-class bias monotonically raises the predicted fake rate on both datasets, but does not improve Macro-F1. On FineFake, the predicted fake rate increases from 55.76% at $\delta = 0$ to 58.89% and 61.85% at $\delta = +0.25$ and $+0.50$, while Macro-F1 decreases from 0.8057 to 0.8033

Table 8: Constant fake-logit bias control on the test sets. Pred. Fake denotes the predicted fake-class proportion.

Dataset	Bias δ	Macro-F1 ↑	Pred. Fake (%)
FineFake	0	0.8057	55.76
	+0.25	0.8033	58.89
	+0.50	0.7971	61.85
Weibo	0	0.9493	52.12
	+0.25	0.9486	52.33
	+0.50	0.9451	52.67

Table 9: Complete ablation studies on Weibo and FineFake. ΔACC is relative to the full ACHER model.

Method	Weibo		FineFake	
	ACC	ΔACC	ACC	ΔACC
ACHER (full model)	**0.949**	–	**0.809**	–
(a) Availability and bidirectional reasoning				
W/o evidence availability	0.942	−0.007	0.804	−0.005
W/o null-hypothesis competition	0.943	−0.006	0.805	−0.004
W/o sparse localization	0.941	−0.008	0.801	−0.008
Shared relation descriptor	0.943	−0.006	0.805	−0.004
Text-to-visual only	0.939	−0.010	0.798	−0.011
Visual-to-text only	0.941	−0.008	0.804	−0.005
(b) Bidirectional decomposition and the four mechanisms				
W/o symmetric–antisymmetric decomposition	0.931	−0.018	0.785	−0.024
W/o compatibility mechanism	0.942	−0.007	0.794	−0.015
W/o conflict-magnitude mechanism	0.938	−0.011	0.801	−0.008
W/o directional-conflict mechanism	0.945	−0.004	0.802	−0.007
W/o unverified mechanism	0.945	−0.004	0.804	−0.005
(c) State constraints and training objective				
W/o evidence-state gating	0.942	−0.007	0.799	−0.010
W/o mechanism-level auxiliary supervision	0.943	−0.006	0.801	−0.008
W/o structural semantic constraint	0.943	−0.006	0.805	−0.004

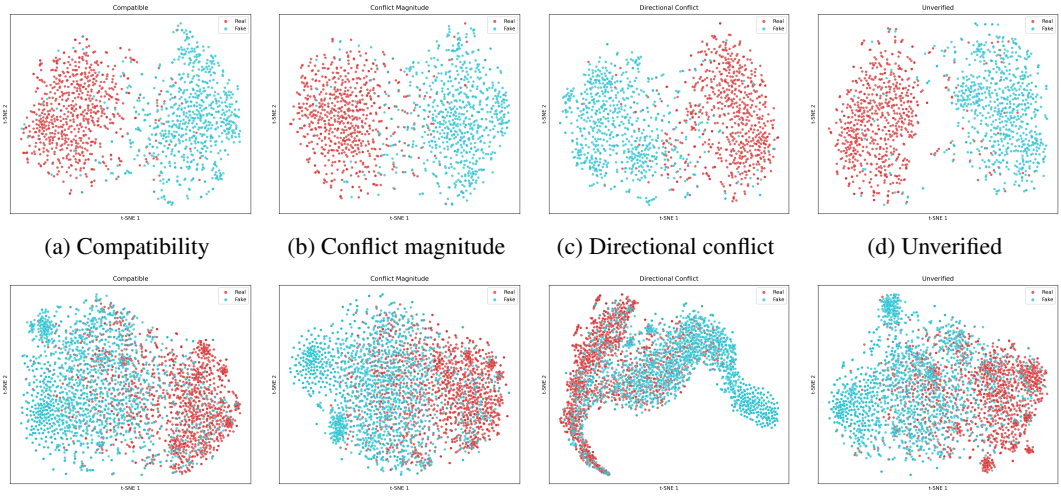
and 0.7971. Weibo exhibits the same trend, with Macro-F1 decreasing from 0.9493 to 0.9486 and 0.9451. Because adding a constant offset to the fake-vs.-real logit margin preserves sample ordering, AUROC and class-wise AP remain unchanged; their test-set values are reported in Table 7. Together with the validation-selected threshold analysis above, these results indicate that the reported performance is stable to the examined operating-point variations and is not explained by a favorable class-specific logit offset.

A.4 Complete Ablation Results

Table 9 reports the complete ablation matrix. The main text retains the variants that directly test the two-level decomposition and the four evidence mechanisms; here we additionally report descriptor sharing and the two training-objective ablations.

A.5 Additional Representation Visualizations

The full structured representations are shown in Fig. 2. Fig. 3 further visualizes the four mechanism-specific embeddings. Their different class-related geometries provide a qualitative view of the mechanism-specific representations, complementing the intervention analyses in the main text.



(a) Compatibility (b) Conflict magnitude (c) Directional conflict (d) Unverified

(e) Compatibility (f) Conflict magnitude (g) Directional conflict (h) Unverified

Figure 3: Mechanism-specific t-SNE visualizations on Weibo (top) and FineFake (bottom).

A.6 Physical-Presence Audit for Mechanistic Diagnostics

Availability is intended to describe whether the *content* of the other modality supplies sufficient counterpart evidence, not whether an input file is physically absent. We verify input presence for all samples used by the evidence-state diagnostics. FineFake has physically present images for all $3{,}376/3{,}376$ validation instances and all $3{,}384/3{,}384$ test instances; Weibo likewise has images for all $843/843$ validation and $1{,}460/1{,}460$ test instances. Text inputs are non-empty after the corresponding preprocessing pipelines. Consequently, the Availability variation studied by deletion, replacement, and regime analyses occurs among physically complete text–image pairs. Thus, low Availability in these analyses is not attributable to physically missing image or text inputs.

A.7 Full Evidence-Deletion Results

The compact main-text table retains the full deletion-dose sequence for ΔA, because the targeted–random contrast and its dose dependence directly test the intended Availability role. Table 10 expands this analysis with the induced Unverified response and state conservation. Since $U = 1 - A$, $\Delta U = -\Delta A$ is expected by construction rather than an independent semantic result; likewise, $\Delta\Sigma = \Delta(S + K + U)$ should remain at numerical precision because the three states are closed by definition.

Effect definitions. For direction $d \in \{T \to V, V \to T\}$ and requested deletion ratio r, let $\mathcal{C}_{d,r}$ be the set of samples for which targeted deletion is valid and let $N_{d,r} = |\mathcal{C}_{d,r}|$. For any direction-level state quantity $X \in \{A, S, K, U\}$, the targeted effect reported in the table is the mean paired change

$$\Delta X^{d,r}_{\text{Tar.}} = \frac{1}{N_{d,r}} \sum_{i \in \mathcal{C}_{d,r}} \left(X^{d,r}_{i,\text{Tar.}} - X^{d}_{i,\text{obs}} \right). \tag{16}$$

Matched-random deletion removes exactly the same number of valid candidates as targeted deletion on the same valid cohort. With $R_{\text{rand}} = 4$ independent random draws, the reported random-control effect averages all paired sample–draw evaluations,

$$\Delta X^{d,r}_{\text{Rand.}} = \frac{1}{R_{\text{rand}} N_{d,r}} \sum_{q=1}^{R_{\text{rand}}} \sum_{i \in \mathcal{C}_{d,r}} \left(X^{d,r,q}_{i,\text{Rand.}} - X^{d}_{i,\text{obs}} \right). \tag{17}$$

For either intervention $\bullet \in \{\text{Tar.}, \text{Rand.}\}$, the reported state-conservation diagnostic is

$$\Delta\Sigma^{d,r}_{\bullet} = \Delta S^{d,r}_{\bullet} + \Delta K^{d,r}_{\bullet} + \Delta U^{d,r}_{\bullet}. \tag{18}$$

The N column therefore counts unique targeted-valid samples, while each matched-random mean pools four evaluations per valid sample.

Table 10: Full targeted evidence-deletion results. "Tar." removes candidates with the highest received grounding mass; "Rand." removes the same number of valid candidates uniformly at random and averages four draws.

Dataset	Direction	Del.	$\Delta A_{\text{Tar.}}$	$\Delta U_{\text{Tar.}}$	$\Delta A_{\text{Rand.}}$	$\Delta U_{\text{Rand.}}$	$\Delta\Sigma_{\text{Tar.}}$	$\Delta\Sigma_{\text{Rand.}}$
Weibo	$T\to V$	5%	-0.0298	$+0.0298$	-0.0005	$+0.0005$	-2.4×10^{-10}	-3.3×10^{-10}
		10%	-0.0465	$+0.0465$	-0.0010	$+0.0010$	-1.3×10^{-10}	-1.9×10^{-10}
		15%	-0.0550	$+0.0550$	-0.0015	$+0.0015$	$+1.9\times10^{-10}$	-2.6×10^{-10}
		20%	-0.0616	$+0.0616$	-0.0022	$+0.0022$	$+4.9\times10^{-10}$	-8.8×10^{-10}
	$V\to T$	5%	-0.0136	$+0.0136$	-0.0010	$+0.0010$	-6.2×10^{-10}	-1.0×10^{-9}
		10%	-0.0206	$+0.0206$	-0.0020	$+0.0020$	-2.2×10^{-9}	-5.7×10^{-10}
		15%	-0.0260	$+0.0260$	-0.0030	$+0.0030$	-7.5×10^{-10}	-9.5×10^{-10}
		20%	-0.0306	$+0.0306$	-0.0042	$+0.0042$	-5.2×10^{-10}	-1.2×10^{-9}
FineFake	$T\to V$	5%	-0.0172	$+0.0172$	-0.0005	$+0.0005$	$+1.2\times10^{-9}$	$+3.3\times10^{-10}$
		10%	-0.0252	$+0.0252$	-0.0010	$+0.0010$	$+8.8\times10^{-12}$	$+3.3\times10^{-10}$
		15%	-0.0295	$+0.0295$	-0.0015	$+0.0015$	-2.6×10^{-10}	$+4.3\times10^{-10}$
		20%	-0.0334	$+0.0334$	-0.0021	$+0.0021$	$+9.5\times10^{-10}$	$+7.8\times10^{-10}$
	$V\to T$	5%	-0.0102	$+0.0102$	-0.0013	$+0.0013$	$+2.4\times10^{-10}$	-6.3×10^{-11}
		10%	-0.0150	$+0.0150$	-0.0021	$+0.0021$	-1.1×10^{-10}	$+4.4\times10^{-10}$
		15%	-0.0194	$+0.0194$	-0.0032	$+0.0032$	-2.1×10^{-10}	-7.6×10^{-11}
		20%	-0.0237	$+0.0237$	-0.0044	$+0.0044$	$+2.3\times10^{-10}$	$+5.3\times10^{-10}$

Dose response and matched-random control. Across both datasets and directions, targeted removal produces a monotonic decrease in Availability from 5% through 20% deletion, whereas matched random removal produces much smaller changes. At the 20% dose, the magnitude ratio between targeted and random deletion is 27.8× for Weibo $T \to V$, 7.3× for Weibo $V \to T$, 15.8× for FineFake $T \to V$, and 5.4× for FineFake $V \to T$. Matched-random deletion controls for representation changes caused by removing the same number of candidates; the substantially larger response under targeted deletion associates the Availability change with localized counterpart evidence.

Prediction-direction check. We also examine how increased Unverified mass relates to the final fake-class probability. On FineFake at 20% targeted deletion, removing text-side counterpart evidence changes $p(\text{Fake})$ by $+0.0130$, whereas removing visual-side counterpart evidence changes it by -0.0230, even though both manipulations decrease Availability and increase Unverified mass. The two directions therefore produce opposite changes in $p(\text{Fake})$, indicating that the effect of Unverified depends on the accompanying state-constrained evidence.

A.8 Availability-Matching Sensitivity

Table 11 reports the full sensitivity analysis for counterpart replacement. "Base" is the direction-specific high-Availability/high-Compatible cohort before matching. The ϵ rows are nested post-hoc subsets satisfying $|A^{\text{cf}} - A^{\text{obs}}| < \epsilon$ from the same replacement inference. For FineFake $T \to V$, all Base-cohort samples already satisfy the $\epsilon = 0.05$ criterion, so the Base and $\epsilon = 0.05$ rows are identical. As ϵ is tightened, residual Availability variation decreases or remains unchanged. Under the strict $\epsilon = 0.01$ control used in the main text, all four dataset–direction pairs exhibit $\Delta D > 0$, $\Delta c > 0$, $\Delta S < 0$, and $\Delta K > 0$; the same sign pattern already holds at $\epsilon = 0.02$.

Counterfactual averaging and matched-cohort definitions. The replacement implementation uses $R_{\text{cf}} = 4$ split-global, content-aware derangements. For each sample i, direction d, and quantity X, the four counterfactual runs are first averaged at the sample level,

$$\bar{X}^{\text{cf}}_{i,d} = \frac{1}{R_{\text{cf}}} \sum_{q=1}^{R_{\text{cf}}} X^{\text{cf},q}_{i,d}. \tag{19}$$

All subsequent matching and effect estimates use this same repeat-averaged counterfactual. Let $q^{50}_{A,d}$ and $q^{75}_{S,d}$ denote the median observed Availability and the 75th percentile of observed Compatibility, respectively, for direction d on the analyzed split. The direction-specific Base cohort is

$$\mathcal{B}_d = \left\{ i : A^{\text{obs}}_{i,d} \geq q^{50}_{A,d},\ S^{\text{obs}}_{i,d} \geq q^{75}_{S,d} \right\}. \tag{20}$$

For matching tolerance ϵ, the nested Availability-controlled cohort is

$$\mathcal{C}_{d,\epsilon} = \left\{ i \in \mathcal{B}_d : \left| \bar{A}^{\text{cf}}_{i,d} - A^{\text{obs}}_{i,d} \right| < \epsilon \right\}. \tag{21}$$

For $X \in \{A, D, c, S, K, U\}$, where D denotes the direction-level Availability-weighted mean of the local discrepancy D^d_i defined in Eq. equation 4, the cohort-level counterfactual effect and the residual Availability shift are

$$\begin{aligned} \Delta X_{d,\epsilon} &= \frac{1}{|\mathcal{C}_{d,\epsilon}|} \sum_{i \in \mathcal{C}_{d,\epsilon}} \left(\bar{X}^{\text{cf}}_{i,d} - X^{\text{obs}}_{i,d} \right), \\ \text{Mean}\,|\Delta A|_{d,\epsilon} &= \frac{1}{|\mathcal{C}_{d,\epsilon}|} \sum_{i \in \mathcal{C}_{d,\epsilon}} \left| \bar{A}^{\text{cf}}_{i,d} - A^{\text{obs}}_{i,d} \right|. \end{aligned} \tag{22}$$

For the Base row, the same definitions apply with $\mathcal{C}_{d,\epsilon}$ replaced by $\mathcal{B}_d$. The different ϵ rows are therefore post-hoc filters of the same counterfactual inference rather than separate model runs.

Robustness to Availability matching. The controlled response is stable across the reported matching tolerances. At $\epsilon = 0.02$, all four dataset–direction pairs already show the same Relation-state response. Tightening the constraint to $\epsilon = 0.01$ reduces Mean $|\Delta A|$ to $0.0039/0.0054$ on Weibo and $0.0046/0.0045$ on FineFake for the text-to-visual and visual-to-text directions, respectively, while the positive D and c responses persist. In the visual-to-text direction, ΔK becomes positive on both datasets once Availability is tightly matched, whereas looser matching permits the decrease in A to dominate $K = Ac$.

Sample-level consistency. The strict-matching effects occur in a majority of samples. Across the four $\epsilon = 0.01$ subsets, D increases for 67.2–83.6% of samples, c increases for 67.7–83.6%, S decreases for 72.7–89.6%, and K increases for 64.4–79.1%. Because $S = A(1 - c)$, $K = Ac$, and $U = 1 - A$, the $S/K/U$ redistribution is a consistency consequence of the hierarchy rather than an independent semantic test. Grounded disagreement D and conditional conflict c therefore increase at the sample level under tightly bounded Availability.

Complementarity of the interventions. Evidence deletion and counterpart replacement target different questions in the same hierarchy. Deletion changes whether enough localized counterpart evidence remains and selectively reduces Availability relative to matched random deletion. Replacement instead changes the relation of the available counterpart to the fixed source; after matching Availability, it changes D and c without routing the response through the Unverified state. Together, the two interventions probe distinct stages of the hierarchy: evidence sufficiency and conditional relation.

A.9 Independent Bootstrap Robustness Details

This subsection reports additional diagnostics for the independent checkpoint-only replacement rerun summarized in Table 12. The rerun uses four validation-global content-aware derangements, excludes same-identity replacements, and applies the same sample-level Availability constraints. The independently generated derangements yield different retained cohort sizes from Table 11.

Strict paired percentile-bootstrap confidence intervals. The primary robustness rerun uses the strict sample-level rule $|A^{\text{cf}} - A^{\text{obs}}| < 0.01$ and $B = 2{,}000$ paired bootstrap resamples. For a retained cohort of size N, let

$$\delta^X_i = \bar{X}^{\text{cf}}_i - X^{\text{obs}}_i, \qquad \widehat{\Delta}_X = \frac{1}{N} \sum_{i=1}^{N} \delta^X_i \tag{23}$$

Table 11: Full Availability-controlled counterpart-replacement sensitivity. Δ denotes counterfactual minus observed. Panel (a) reports Availability matching and relation responses; panel (b) reports the induced state redistribution.

(a) Availability matching and relation response

Dataset	Direction	Cohort	N	Mean $\lvert\Delta A\rvert$	ΔA	ΔD	Δc
Weibo	$T\to V$	Base	254	0.0064	-0.0021	$+0.0045$	$+0.0066$
		$\epsilon = 0.05$	253	0.0062	-0.0019	$+0.0046$	$+0.0068$
		$\epsilon = 0.02$	245	0.0055	-0.0011	$+0.0048$	$+0.0070$
		$\epsilon = 0.01$	**205**	**0.0039**	$+0.0002$	$+0.0047$	$+0.0068$
	$V\to T$	Base	352	0.0246	-0.0236	$+0.0042$	$+0.0048$
		$\epsilon = 0.05$	339	0.0234	-0.0223	$+0.0043$	$+0.0049$
		$\epsilon = 0.02$	146	0.0106	-0.0086	$+0.0063$	$+0.0072$
		$\epsilon = 0.01$	**67**	**0.0054**	-0.0025	$+0.0075$	$+0.0086$
FineFake	$T\to V$	Base	699	0.0101	-0.0070	$+0.0045$	$+0.0063$
		$\epsilon = 0.05$	699	0.0101	-0.0070	$+0.0045$	$+0.0063$
		$\epsilon = 0.02$	610	0.0078	-0.0045	$+0.0042$	$+0.0059$
		$\epsilon = 0.01$	**402**	**0.0046**	-0.0017	$+0.0035$	$+0.0049$
	$V\to T$	Base	659	0.0444	-0.0421	-0.0083	-0.0112
		$\epsilon = 0.05$	376	0.0262	-0.0222	$+0.0007$	$+0.0013$
		$\epsilon = 0.02$	141	0.0102	-0.0044	$+0.0107$	$+0.0151$
		$\epsilon = 0.01$	**67**	**0.0045**	-0.0013	$+0.0128$	$+0.0180$

(b) State redistribution

Dataset	Direction	Cohort	ΔS	ΔK	ΔU
Weibo	$T\to V$	Base	-0.0055	$+0.0034$	$+0.0021$
		$\epsilon = 0.05$	-0.0055	$+0.0036$	$+0.0019$
		$\epsilon = 0.02$	-0.0053	$+0.0043$	$+0.0011$
		$\epsilon = 0.01$	-0.0046	$+0.0048$	-0.0002
	$V\to T$	Base	-0.0200	-0.0036	$+0.0236$
		$\epsilon = 0.05$	-0.0192	-0.0032	$+0.0223$
		$\epsilon = 0.02$	-0.0107	$+0.0021$	$+0.0086$
		$\epsilon = 0.01$	-0.0071	$+0.0046$	$+0.0025$
FineFake	$T\to V$	Base	-0.0087	$+0.0017$	$+0.0070$
		$\epsilon = 0.05$	-0.0087	$+0.0017$	$+0.0070$
		$\epsilon = 0.02$	-0.0069	$+0.0024$	$+0.0045$
		$\epsilon = 0.01$	-0.0044	$+0.0028$	$+0.0017$
	$V\to T$	Base	-0.0194	-0.0227	$+0.0421$
		$\epsilon = 0.05$	-0.0150	-0.0072	$+0.0222$
		$\epsilon = 0.02$	-0.0128	$+0.0084$	$+0.0044$
		$\epsilon = 0.01$	-0.0127	$+0.0114$	$+0.0013$

be the paired sample effect and its mean. For bootstrap replicate b, indices $I_1^{(b)}, \ldots, I_N^{(b)}$ are sampled with replacement from $\{1, \ldots, N\}$ and

$$\widehat{\Delta}_X^{*(b)} = \frac{1}{N} \sum_{j=1}^{N} \delta^X_{I_j^{(b)}}. \tag{24}$$

The reported 95% interval is the percentile interval

$$\mathrm{CI}_{95\%}(\Delta X) = \left[Q_{0.025}\left(\{\widehat{\Delta}_X^{*(b)}\}_{b=1}^{B} \right), Q_{0.975}\left(\{\widehat{\Delta}_X^{*(b)}\}_{b=1}^{B} \right) \right]. \tag{25}$$

Table 12: Independent strict-matching replacement rerun with paired percentile-bootstrap 95% CIs (2,000 resamples). Δ denotes counterfactual minus observed. Each effect estimate is listed with its corresponding 95% CI.

Dataset	Direction	N	Mean $\lvert\Delta A\rvert$	Effect	Estimate	95% CI
Weibo	$T \to V$	106	0.0038	ΔD	$+0.0056$	$[0.0043, 0.0069]$
				Δc	$+0.0056$	$[0.0043, 0.0069]$
				ΔS	-0.0055	$[-0.0066, -0.0043]$
				ΔK	$+0.0058$	$[0.0043, 0.0074]$
	$V \to T$	48	0.0049	ΔD	$+0.0056$	$[0.0036, 0.0076]$
				Δc	$+0.0056$	$[0.0035, 0.0076]$
				ΔS	-0.0058	$[-0.0076, -0.0040]$
				ΔK	$+0.0033$	$[0.0018, 0.0047]$
FineFake	$T \to V$	399	0.0047	ΔD	$+0.0036$	$[0.0028, 0.0043]$
				Δc	$+0.0036$	$[0.0028, 0.0043]$
				ΔS	-0.0045	$[-0.0053, -0.0037]$
				ΔK	$+0.0028$	$[0.0020, 0.0035]$
	$V \to T$	55	0.0045	ΔD	$+0.0167$	$[0.0132, 0.0201]$
				Δc	$+0.0167$	$[0.0132, 0.0203]$
				ΔS	-0.0160	$[-0.0190, -0.0130]$
				ΔK	$+0.0147$	$[0.0117, 0.0176]$

Table 13: Independent replacement-rerun sensitivity to the Availability-matching tolerance.

Dataset	Direction	Match	N	Mean $\lvert\Delta A\rvert$	ΔA	ΔD	ΔS	ΔK
Weibo	$T \to V$	$\epsilon = 0.02$	126	0.0054	-0.0004	$+0.0053$	-0.0056	$+0.0052$
	$T \to V$	$\epsilon = 0.01$	106	0.0038	$+0.0004$	$+0.0056$	-0.0055	$+0.0058$
	$V \to T$	$\epsilon = 0.02$	104	0.0103	-0.0083	$+0.0067$	-0.0107	$+0.0025$
	$V \to T$	$\epsilon = 0.01$	48	0.0049	-0.0025	$+0.0056$	-0.0058	$+0.0033$
FineFake	$T \to V$	$\epsilon = 0.02$	633	0.0083	-0.0047	$+0.0038$	-0.0066	$+0.0019$
	$T \to V$	$\epsilon = 0.01$	399	0.0047	-0.0017	$+0.0036$	-0.0045	$+0.0028$
	$V \to T$	$\epsilon = 0.02$	111	0.0102	-0.0055	$+0.0108$	-0.0135	$+0.0080$
	$V \to T$	$\epsilon = 0.01$	55	0.0045	-0.0013	$+0.0167$	-0.0160	$+0.0147$

Thus, the resampling unit is the paired sample-level counterfactual effect rather than independently resampled observed and counterfactual sets. Table 12 reports the resulting uncertainty. All four dataset–direction pairs retain positive ΔD and Δc with 95% CIs strictly above zero; ΔS is negative and ΔK positive with intervals excluding zero. Mean $|\Delta A|$ remains below 0.005 in every retained strict cohort.

Sensitivity to the matching tolerance. The expected sign pattern is already present under $\epsilon = 0.02$ and remains under the stricter $\epsilon = 0.01$ control. Table 13 reports the corresponding point estimates.

Sample-level consistency. Under the strict $\epsilon = 0.01$ control, grounded disagreement D increases for 73.6% and 81.3% of Weibo samples in the $T \to V$ and $V \to T$ directions, with 95% Wilson intervals $[0.645, 0.810]$ and $[0.681, 0.898]$. On FineFake, the corresponding rates are 70.2% and 92.7%, with intervals $[0.655, 0.745]$ and $[0.827, 0.971]$. Compatible-state mass decreases in 79.2%/81.3% of Weibo and 71.7%/92.7% of FineFake samples, while Conflict-state mass increases in 72.6%/81.3% and 66.4%/92.7%, respectively. Since $S = A(1 - c)$ and $K = Ac$, the state-mass responses are consistency consequences of the hierarchy; the nontrivial statistical evidence is the positive D/c response under tightly bounded Availability.

Table 14: Validation-derived median Availability thresholds used for the primary evidence-regime partition.

Dataset	m_T ($T \to V$)	m_V ($V \to T$)
FineFake	0.6792	0.6155
Weibo	0.6737	0.6034

A.10 Evidence-Regime Analysis Details

Protocol and regime construction. The evidence-regime analysis is performed with the selected full ACHER checkpoint and introduces no additional training. For each dataset, we first run the frozen model on the validation split and compute the median direction-level Availability in the text-to-visual and visual-to-text directions. These medians are then frozen and applied to the test set. Let m_T and m_V denote the validation medians for A_T and A_V, respectively. The four primary regimes are

$$\mathcal{R}_{HH} = \{A_T \geq m_T,\ A_V \geq m_V\}, \quad \mathcal{R}_{HL} = \{A_T \geq m_T,\ A_V < m_V\},$$
$$\mathcal{R}_{LH} = \{A_T < m_T,\ A_V \geq m_V\}, \quad \mathcal{R}_{LL} = \{A_T < m_T,\ A_V < m_V\},$$

and the asymmetric cohort is $\mathcal{R}_{HL} \cup \mathcal{R}_{LH}$. The resulting thresholds are shown in Table 14. The thresholds are computed on validation data and then fixed for test-set grouping.

Fixed-cohort mechanism knockouts. For each frozen test cohort, we perform four separate inference-time interventions by setting one learned mechanism representation to zero immediately before the classifier: Compatibility, Conflict Magnitude, Directional Conflict, or Unverified. No model parameter is updated and the cohort membership remains identical across the full model and all knockout variants. Writing $m \in \{S, K, D, U\}$ for Compatibility, Conflict Magnitude, Directional Conflict, and Unverified, respectively, the post-hoc prediction is

$$\hat{y}_i^{\neg m} = C\big(\boldsymbol{z}_i^S, \ldots, \mathbf{0}_m, \ldots, \boldsymbol{z}_i^U\big), \tag{26}$$

where C is the frozen final classifier and all non-target mechanism representations are left unchanged. For a fixed regime $\mathcal{R}_r$, Table 15 reports the Macro-F1 change in percentage points as

$$\Delta_{\neg m}^{(r)} = 100\left[\mathrm{MacroF1}(\mathcal{R}_r; \hat{y}^{\neg m}) - \mathrm{MacroF1}\big(\mathcal{R}_r; \hat{y}^{\mathrm{full}}\big)\right]. \tag{27}$$

Thus, negative values indicate that removing the mechanism reduces Macro-F1 within the same frozen cohort. The inference-time knockouts evaluate reliance on learned mechanisms within fixed cohorts, whereas the architectural ablations in Table 9 evaluate component contributions during training.

Because the mechanisms are jointly learned and the knockout is applied only after training, occasional positive subgroup-level Δ values can arise from post-hoc interactions among the remaining learned representations and do not by themselves imply a full-dataset benefit from removal. Consistent with this distinction, the full-test-set rows show no positive gain from removing any mechanism on either dataset.

The regime analysis specifically tests the Unverified pathway, whose strength is directly tied to evidence availability. Its knockout effect becomes progressively more negative from Both-high to Asymmetric to Both-low Availability on both datasets, changing from $+0.06$ to -0.11 to -0.52 on FineFake and from -0.45 to -0.92 to -2.35 on Weibo.

Interpretation. The Unverified knockout effect becomes more negative from both-high to asymmetric to both-low Availability on both datasets: $+0.06/-0.11/-0.52$ points on FineFake and $-0.45/-0.92/-2.35$ points on Weibo. This pattern links model-estimated Availability regimes to final prediction behavior: the trained classifier depends more strongly on Unverified reasoning in cohorts with lower estimated counterpart-evidence availability. Directional Conflict is gated by $G_K = |K_T - K_V|$, which measures conflict-state asymmetry rather than Availability asymmetry.

The raw Macro-F1 values across regimes should also be interpreted cautiously because the class mixtures differ strongly across cohorts (e.g., FineFake ranges from 28.55% fake in Both-high to

Table 15: Complete primary evidence-regime analysis. Δ values are Macro-F1 changes in percentage points for inference-time mechanism knockout minus the full checkpoint. "Asymmetric" is the union of the two off-diagonal Availability regimes.

Dataset	Regime	N	Fake (%)	Full F1	$\Delta_{\neg S}$	$\Delta_{\neg K}$	$\Delta_{\neg D}$	$\Delta_{\neg U}$
FineFake	All	3,384	56.35	0.8057	-0.75	-0.32	-1.00	-0.11
	Both-high	802	28.55	0.7350	$+0.51$	$+0.08$	-1.15	$+0.06$
	$T \to V$ high / $V \to T$ low	845	83.08	0.8035	-2.73	-1.26	-0.62	-0.27
	$T \to V$ low / $V \to T$ high	879	39.70	0.7511	-1.28	-0.45	-0.18	-0.48
	Both-low	858	73.08	0.6520	-3.19	$+0.18$	-1.69	-0.52
	Asymmetric	1,724	60.96	0.8191	-1.41	-0.78	-0.20	-0.11
Weibo	All	1,460	51.44	0.9493	-0.07	-0.34	-0.27	-1.10
	Both-high	287	66.55	0.9446	-0.03	$+0.00$	-1.13	-0.45
	$T \to V$ high / $V \to T$ low	440	25.68	0.9436	-0.21	-0.63	$+0.00$	-1.50
	$T \to V$ low / $V \to T$ high	435	68.05	0.9603	-0.03	-0.55	$+0.00$	-0.79
	Both-low	298	50.67	0.9127	-0.01	-0.33	-0.33	-2.35
	Asymmetric	875	46.74	0.9610	-0.11	-0.46	$+0.00$	-0.92

73.08% in Both-low). We therefore use within-cohort knockout contrasts rather than cross-cohort raw Macro-F1 differences for regime-level interpretation. Together with the trained ablations, targeted evidence deletion, and Availability-controlled counterpart replacement, this analysis adds classifier-level evidence that the Unverified pathway is used more strongly in lower-Availability regimes.