TRUTHDEBATE: MULTI-AGENT COLLABORATIVE REASONING FOR DEEPFAKE NEWS DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

While multimodal large language models (MLLMs) have made significant progress for deepfake detection, manipulation of deepfake often exhibits various semantically coherent yet deceptive content such as face swap, face emotion editing or text fabrication, making single-view analysis of one MLLM insufficient. LLM-based multi-agent systems (MAS) provide a promising way that collectively mine deceptive content from complementary perspectives to combat this threat. However, conventionally MAS suffer from high computational costs and significant optimization difficulties. To solve these challenges, we propose an adaptive multi-agent framework named TruthDebate for deepfake detection. TruthDebate consists of two core stages: (1) we design a multimodal clue extraction model to capture visual and textual clues of deceptive content to provide prior knowledge as the input of the MAS; (2) Guided by these clues, we design an efficiency-aware multi-agent collaborative reasoning mechanism that includes multimodal clue analysts, evidence integrators, an evidence auditor, and a decision-maker for cooperative reasoning to adaptively optimize MAS topology and routing-based inference. Extensive experiments on the MLLM-Driven synthetic multimodal deepfake dataset demonstrate the effectiveness of our framework. Our framework significantly outperforms multimodal detectors and multi-agent methods while reducing token consumption. Code is available at https://anonymous.4open.science/r/TruthDebate-3131/.

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have created new opportunities for deepfake detection. At the same time, generative models make it increasingly easy to produce deceptive news content with plausible images and texts that remain semantically consistent across modalities Zhang et al. (2026). Such content may involve face swap, face emotion editing, or text fabrication, while remaining semantically coherent at a surface level. Detecting these manipulations requires jointly examining visual traces, facial information, textual claims, and image–text relations.

Existing multimodal detectors have progressed from unimodal artifact detection to modeling cross-modal consistency and manipulation-aware representations Shao et al. (2023; 2024); Ma et al. (2024); Li et al. (2025b). Recent MLLM-based methods further use visual-language reasoning for misinformation detection and explanation Qi et al. (2024); Liu et al. (2024); Huang et al. (2025b). However, these methods generally rely on a single model with a fixed inference process. A single-view analysis may overlook clues that are distributed across different modalities or manipulation types. Moreover, a prediction score alone provides limited support for examining how different pieces of evidence contribute to the final decision.

LLM-based multi-agent systems (MAS) offer a natural way to address this limitation by assigning complementary analysis roles and integrating their outputs. Recent studies have explored sequential collaboration, graph-based agent organization, role differentiation, and coordinated problem solving Zhang et al. (2024d); Zhuge et al. (2024); Li et al. (2025a); Klein et al. (2025). Nevertheless, directly applying a complete MAS to every deepfake sample is inefficient: multiple agent calls and repeated reports substantially increase language-model token consumption. Further, unreliable or redundant reports can affect the quality of cooperative reasoning, making the agent structure difficult to optimize in practice Huang et al. (2025a); Zhu et al. (2025a).

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.

We propose TRUTHDEBATE, an adaptive multi-agent framework for deepfake detection. TRUTHDEBATE consists of two stages. The first stage, *Multimodal Clue Extraction*, captures visual and textual clues of deceptive content and provides prior knowledge for the MAS. It inserts learnable multimodal clue embeddings between visual and textual representations and uses complementary reasoning branches to predict scores for face swap, face emotion editing, and text fabrication. These three scores form a three-dimensional prior.

The second stage, *Efficiency-aware Multi-Agent Collaborative Reasoning*, uses the prior to determine whether a sample requires further cooperative analysis. Samples with sufficiently confident prior scores are directly predicted, whereas the remaining samples are routed to multimodal clue analysts. Their reports are processed by evidence integrators and assessed by an independent evidence auditor before a decision-maker produces the final prediction with supporting evidence. During training, prediction feedback is used to estimate role reliability, construct corrective cues, and select reliable analysts, thereby optimizing the MAS topology. These quantities are fixed during testing, where routing-based inference avoids unnecessary agent calls.

We evaluate TRUTHDEBATE on the MLLM-Driven Synthetic Multimodal (MDSM) dataset Zhang et al. (2026) across five news domains. The results show that TRUTHDEBATE outperforms supervised multimodal detectors and multi-agent baselines while reducing language-model token consumption.

Our contributions are summarized as follows:

- We propose TRUTHDEBATE, an adaptive framework that combines multimodal clue extraction with efficiency-aware multi-agent collaborative reasoning for deepfake detection.
- We design a multimodal clue extraction model that uses learnable clue embeddings and complementary reasoning branches to produce a three-dimensional prior over face swap, face emotion editing, and text fabrication.
- We develop an efficiency-aware collaborative reasoning mechanism that combines feedback-guided MAS optimization with routing-based inference, improving cooperative evidence analysis while reducing unnecessary language-model token consumption.

2 RELATED WORK

2.1 DEEPFAKE DETECTION

Deepfake and multimodal fake news detection has evolved from combining visual and textual features to modeling fine-grained manipulation cues and cross-modal inconsistencies. Early studies learn event-invariant multimodal representations or establish fine-grained multimodal benchmarks Wang et al. (2018); Nakamura et al. (2020). Subsequent methods explore cross-modal ambiguity, semantic interaction, and consistency reasoning Chen et al. (2022); Wu et al. (2023). HAMMER and HAMMER++ Shao et al. (2023; 2024) learn manipulation-aware representations for classification and grounding, while Event-Radar Ma et al. (2024) models event-level multimodal inconsistencies.

Recent work further incorporates external knowledge, domain information, and fine-grained manipulation attributes. AKA-Fake Zhang et al. (2024c) selects informative knowledge for multimodal detection, while ERIC-FND and DAMMFND Cao et al. (2025); Lu et al. (2025) improve external-information utilization and domain-aware multimodal modeling, respectively. AMG further introduces multi-granularity fake-news attribution Guo et al. (2025a).

Large multimodal models have also been applied to misinformation detection and explanation. SNIFFER Qi et al. (2024) and FKA-Owl Liu et al. (2024) examine image–text relations with multimodal language models, whereas SIDA and M2F2-Det Huang et al. (2025b); Guo et al. (2025b) address detection, localization, or explanation. CSCL Li et al. (2025b) improves intra- and cross-modal consistency, while SDML Zhu et al. (2025b) introduces a multimodal fake-news dataset and a shallow-deep multitask framework. The MDSM dataset Zhang et al. (2026) focuses on semantically coherent image–text manipulations generated with MLLMs, and Forensics-Bench Wang et al. (2025) evaluates large vision–language models across diverse forgery tasks. In contrast, TRUTHDE-

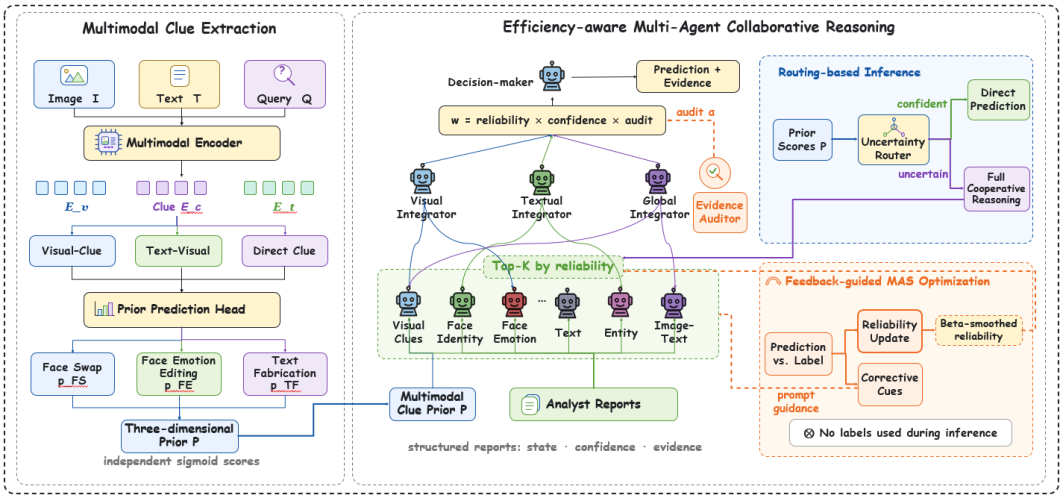

Figure 1: Overall pipeline of TRUTHDEBATE. Multimodal Clue Extraction captures visual and textual clues and produces a three-dimensional prior. Efficiency-aware Multi-Agent Collaborative Reasoning uses this prior for cooperative evidence integration and decision, training-time feedback-guided MAS optimization, and routing-based inference.

BATE uses dimension-specific multimodal clues to guide adaptive multi-agent collaborative reasoning.

2.2 MULTI-AGENT COLLABORATIVE REASONING

LLM-based multi-agent systems improve complex reasoning through role specialization and information exchange. MetaGPT Hong et al. (2024) organizes agents with standardized role-specific workflows, whereas Chain-of-Agents (CoA) Zhang et al. (2024d) aggregates information through sequential collaboration. MAGIS Tao et al. (2024) and SMART Yue et al. (2025) coordinate specialized agents for collaborative problem solving. Multi-persona self-collaboration has also been explored to improve reasoning through dynamically assigned personas Wang et al. (2024).

Recent studies investigate adaptive coordination and communication design. ProAgent Zhang et al. (2024a) adapts cooperative behaviors to teammates, GPTSwarm Zhuge et al. (2024) represents agent workflows as optimizable graphs, and role-differentiated debate further improves collaborative reasoning Li et al. (2025a). Co-LLM learns when a base model should invoke assistant models during decoding Shen et al. (2024), while MachineSoM analyzes collaboration mechanisms through a social-psychology perspective Zhang et al. (2024b).

However, additional agents and communication rounds increase inference cost and can propagate unreliable information. MAGDi Chen et al. (2024) distills multi-agent interaction graphs, whereas Fleet of Agents Klein et al. (2025) studies the balance between reasoning quality and computational cost. Recent work further examines collaboration with faulty agents Huang et al. (2025a) and evaluates different communication structures Zhu et al. (2025a). Unlike these general-purpose systems, TRUTHDEBATE uses multimodal clue priors, feedback-derived analyst selection, evidence auditing, and confidence-based routing for efficient deepfake detection.

3 METHODOLOGY

We present TRUTHDEBATE, which consists of two stages: multimodal clue extraction and efficiency-aware multi-agent collaborative reasoning. The overall framework is illustrated in Figure 1.

3.1 MULTIMODAL CLUE EXTRACTION

Deepfake news may contain semantically coherent yet deceptive features, including face swap, face emotion editing, and text fabrication. To capture visual and textual clues of deceptive content, we design a multimodal clue extraction model that introduces learnable multimodal clue embeddings between visual and textual representations. Three complementary reasoning branches then extract task-relevant clues and produce prediction scores for the three manipulation types. These scores provide prior knowledge as input to the MAS.

3.1.1 LEARNABLE MULTIMODAL CLUE EMBEDDINGS

To adapt the pretrained multimodal encoder to deepfake news detection while preserving its pretrained knowledge, we introduce a set of learnable multimodal clue embeddings to capture visual and textual clues of deceptive content. Formally, the clue embeddings are denoted as $E_c \in \mathbb{R}^{N_c \times d}$, where N_c is the number of clue embeddings and d is the embedding dimension. The textual input is processed by a text embedding layer to obtain $E_t \in \mathbb{R}^{N_t \times d}$, while the news image is encoded by the vision backbone and projected into visual embeddings $E_v \in \mathbb{R}^{N_v \times d}$. The visual, clue, and textual embeddings are concatenated to form the multimodal input sequence $S_{\text{in}} = [E_v; E_c; E_t]$, where $[\cdot\,;\cdot]$ denotes concatenation along the token dimension. The resulting sequence is subsequently processed by three complementary reasoning branches.

3.1.2 MULTI-BRANCH CLUE REASONING AND PRIOR PREDICTION

The multimodal input sequence is processed by three complementary reasoning branches. First, the Visual Clue Interaction Branch models the interaction between visual embeddings and multimodal clue embeddings, producing the representation r_{vc}. Second, the Text–Visual Consistency Branch uses textual embeddings to attend to visual and clue features, producing r_{tv} for image–text consistency analysis. Third, the Direct Clue Branch aggregates the multimodal clue embeddings into r_c, retaining task-relevant clues directly captured by these embeddings.

The three representations are concatenated and mapped to three prediction scores:

$$\mathbf{p} = \sigma\left(g_\phi\left([r_{vc}; r_{tv}; r_c]\right)\right), \tag{1}$$

where g_ϕ is a learnable fusion head and $\sigma(\cdot)$ is applied element-wise. The resulting vector $\mathbf{p} \in [0,1]^3$ contains the prediction scores for face swap, face emotion editing, and text fabrication, respectively. They provide prior knowledge as input to the MAS.

3.1.3 TRAINING OBJECTIVE

Training follows a two-stage procedure. We first optimize the multimodal clue embeddings and prediction heads while freezing the pretrained backbone, and then fine-tune the complete model end to end. The overall objective consists of a three-dimensional prediction loss, an auxiliary authenticity loss, and a clue-embedding regularization loss:

$$\mathcal{L} = \lambda_{\text{dim}}\mathcal{L}_{\text{dim}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{2}$$

where the three coefficients control the contributions of the corresponding objectives.

The primary objective is a class-balanced binary cross-entropy loss over the three manipulation types:

$$\mathcal{L}_{\text{dim}} = -\frac{1}{3}\sum_{k=1}^{3}\left[w_k y_k \log p_k + (1-y_k)\log(1-p_k)\right], \tag{3}$$

where $y_k \in \{0,1\}$ is the target for manipulation type k, p_k is its prediction score, and w_k is the corresponding positive-class weight. The three dimensions represent face swap, face emotion editing, and text fabrication, respectively.

We further attach binary prediction heads to the visual, textual, and multimodal clue representations. Their auxiliary authenticity loss is defined as

$$\mathcal{L}_{\text{aux}} = \frac{1}{3}\sum_{m\in\{v,t,c\}} \text{CE}(q_m, b), \tag{4}$$

where q_m is the binary prediction produced from representation m and $b \in \{0, 1\}$ is the real/fake target.

Finally, $\mathcal{L}_{\text{reg}}$ combines orthogonality and feature-magnitude regularization for the multimodal clue embeddings. It reduces redundancy among clue embeddings and prevents their feature energy from concentrating on only a few dimensions. Detailed definitions of the two regularization terms are provided in the appendix.

3.2 Efficiency-aware Multi-Agent Collaborative Reasoning

Guided by the extracted visual and textual clues, this stage coordinates multimodal clue analysts, evidence integrators, an evidence auditor, and a decision-maker. It optimizes the MAS topology using training feedback and reduces unnecessary agent calls through routing-based inference.

3.2.1 Cooperative Evidence Integration and Decision

Evidence integration. The three-dimensional clue prior is first converted into a deterministic dimension-wise report. This report and the reports produced by the selected multimodal clue analysts are processed by visual, textual, and global evidence integrators. The visual and textual integrators aggregate the corresponding modality-specific evidence, while the global integrator reconciles evidence across manipulation types. Each integrator h produces a dimension-wise assessment and a confidence score $q_{h,d}^{(i)}$ for sample i.

Evidence auditing. Alongside the evidence integrators, an independent evidence auditor examines the analyst and integrator reports for missing, duplicated, or conflicting evidence. It assigns an adjustment factor $\alpha_h^{(i)} \in [0, 1]$ to each integrator. Combining this factor with the feedback-derived role reliability $r_{h,d}$, the effective contribution of integrator h is

$$w_{h,d}^{(i)} = r_{h,d}\, q_{h,d}^{(i)}\, \alpha_h^{(i)}. \tag{5}$$

Final decision. The decision-maker combines the integrated reports using their effective weights and the audit results. Reports supported by reliable roles, strong current evidence, and coherent evidence chains receive greater influence in producing the final prediction and its supporting evidence.

3.2.2 Feedback-guided MAS Topology Optimization

The quality of different analysts and evidence integrators may vary across manipulation dimensions. During training, TruthDebate therefore uses prediction feedback to estimate role-specific reliability and adapt the active analyst set. Here, MAS topology refers to the active subset of analysts retained for cooperative reasoning. The ground-truth label is accessed only after the final prediction has been produced; it is not included in any agent prompt.

For each participating role u and manipulation dimension d, let $n_{u,d}$ be the number of its definite judgments on error-triggered feedback samples, and let $c_{u,d}$ be the number of correct judgments. We estimate its reliability using a Beta-smoothed score:

$$r_{u,d} = \frac{c_{u,d} + 2}{n_{u,d} + 4}. \tag{6}$$

Reports marked as `uncertain` are excluded because they do not provide a definite dimension-wise decision. The symmetric prior avoids extreme reliability estimates when only a small number of feedback observations are available.

Let $\mathcal{A}_{\text{TF}}$ denote the candidate text-related analysts. During each training epoch, all candidates are evaluated, and the K analysts with the highest reliability on text fabrication are retained:

$$\mathcal{A}'_{\text{TF}} = \text{TopK}_{a \in \mathcal{A}_{\text{TF}}} \left(r_{a,\text{TF}}; K \right). \tag{7}$$

The face-swap and face-emotion analysts are always retained because they provide dedicated evidence for the two visual manipulation types. For analysts involved in an incorrect decision, we additionally update a concise corrective cue that summarizes the observed failure pattern. After training, the reliability estimates, corrective cues, and selected analyst set are frozen for routing-based test-time inference.

3.2.3 ROUTING-BASED INFERENCE

Routing-based inference determines whether a sample requires full cooperative reasoning. Let $\mathcal{D} = \{\mathrm{FS}, \mathrm{FE}, \mathrm{TF}\}$ denote face swap, face emotion editing, and text fabrication, respectively. For dimension d, the prior prediction is $\hat{y}_d^{(i)} = \mathbb{I}[p_d^{(i)} \geq \theta_d]$, where θ_d is the corresponding decision threshold.

We measure prior confidence by the normalized distance between the prediction score and its decision threshold:

$$\gamma_d^{(i)} = \begin{cases} \dfrac{1}{2} + \dfrac{p_d^{(i)} - \theta_d}{2(1-\theta_d)}, & p_d^{(i)} \geq \theta_d, \\ \dfrac{1}{2} + \dfrac{\theta_d - p_d^{(i)}}{2\theta_d}, & p_d^{(i)} < \theta_d. \end{cases} \tag{8}$$

Thus, a prediction score close to its decision threshold has low confidence, whereas a score close to either endpoint has high confidence.

The inference route is selected as

$$\mathrm{Route}(i) = \begin{cases} \text{direct prior prediction}, & \min\limits_{d \in \mathcal{D}} \gamma_d^{(i)} \geq \tau, \\ \text{full cooperative reasoning}, & \min\limits_{d \in \mathcal{D}} \gamma_d^{(i)} < \tau, \end{cases} \tag{9}$$

where τ is the routing threshold. Samples with sufficiently confident prior predictions bypass the MAS, whereas the remaining samples are processed by the multimodal clue analysts, evidence integrators, evidence auditor, and decision-maker. The decision and routing thresholds are selected on the validation set and fixed during testing. The feedback-derived reliability estimates, corrective cues, and active analyst set are also fixed before testing and remain unchanged throughout test-time inference.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and evaluation protocol. We evaluate TRUTHDEBATE under the sampled setting of MDSM, an MLLM-driven synthetic multimodal deepfake dataset containing paired news images and texts from five domains. Models are trained on the NYT split and evaluated on NYT and four unseen domains: *Guardian*, *USA Today*, *Washington Post*, and *BBC*. Their test sets contain 1,000, 5,647, 1,606, 1,283, and 1,464 samples, respectively. The NYT validation split is used for model selection and for determining the decision and routing thresholds. All thresholds and adaptive statistics are fixed before testing.

Compared methods. We compare TRUTHDEBATE with three categories of baselines: (i) zero-shot multimodal models, including DeepSeek-VL2-27B, Qwen2-VL-72B-Instruct, Qwen3.6-27B, Qwen3-VL-32B-Instruct, and Qwen3.5-122B-A10B; (ii) supervised multimodal detectors, including ViLT, HAMMER, HAMMER++, FKA-Owl, and AMD; and (iii) multi-agent methods, including Chain, Star, Tree, Graph, and DyLAN. The supervised methods are trained on the same NYT split. All multi-agent baselines use the same underlying language model, prior knowledge, input evidence, candidate roles, and decoding configuration as TRUTHDEBATE.

Implementation details. The multimodal clue extraction model is trained on the NYT split to capture visual and textual clues of deceptive content and provide prior knowledge as input to the MAS. We use Qwen3.6-27B as the underlying language model for TRUTHDEBATE and all multi-agent baselines. These methods share the same data splits, prior knowledge, input evidence, candidate roles, and decoding configuration. All experiments are conducted on NVIDIA A100 GPUs. Additional training settings, agent configurations, and prompts are provided in the appendix.

Evaluation metrics. We report ACC and F1 for binary real/fake detection, treating manipulated samples as the positive class. Macro-F1, abbreviated as MaF1 in the tables, measures balanced

Table 1: Performance comparison of TRUTHDEBATE against zero-shot, supervised, and multi-agent baselines on the NYT, Guardian, and USA Today domains. The best results are highlighted in bold.

Setting	Method	NYT				Guardian				USA			
		ACC	F1	MaF1	mAP	ACC	F1	MaF1	mAP	ACC	F1	MaF1	mAP
Zero-shot	DeepSeek-VL2-27B	51.00	1.61	11.23	20.31	47.87	1.80	10.75	21.18	18.93	1.21	5.21	40.96
	Qwen2-VL-72B-Instruct	56.00	20.29	16.87	25.30	52.42	19.19	13.00	22.89	25.47	16.24	7.57	43.42
	Qwen3.6-27B	65.30	49.64	19.39	26.86	63.56	52.12	17.57	26.19	46.64	52.47	12.92	43.58
	Qwen3-VL-32B-Instruct	68.10	61.52	19.82	28.49	62.55	61.01	17.36	25.55	49.63	58.58	12.69	43.23
	Qwen3.5-122B-A10B	63.60	49.58	15.42	21.84	59.77	52.51	13.56	21.62	45.70	52.45	9.81	40.62
Supervised Models	ViLT	91.40	91.02	84.43	77.94	61.75	71.79	31.02	42.38	81.63	89.80	29.32	55.48
	HAMMER	93.60	93.81	82.08	90.98	60.17	71.03	28.72	51.61	82.88	90.46	29.29	64.07
	HAMMER++	94.10	94.05	84.59	89.63	63.31	71.32	30.12	55.51	83.25	90.56	30.60	**67.21**
	FKA-Owl	93.80	93.35	78.39	88.78	77.05	72.74	43.72	63.30	84.12	90.41	33.77	57.93
	AMD	92.30	92.10	78.09	78.71	77.83	74.89	44.35	56.84	82.81	88.89	43.27	58.48
Multi-Agent Baselines	Chain	90.30	90.82	83.44	85.11	74.64	73.29	49.43	57.72	85.49	91.30	51.97	64.85
	Star	83.60	85.61	79.45	81.22	71.72	72.87	44.79	51.25	85.31	91.28	53.27	63.42
	Tree	77.80	81.41	69.61	76.56	68.60	71.64	41.39	49.00	85.55	91.44	52.39	63.94
	Graph	76.20	80.33	66.53	75.61	67.13	70.14	40.68	52.29	85.31	91.30	49.54	61.34
	DyLAN	90.20	90.79	84.28	83.83	74.50	73.97	47.51	53.39	85.80	91.67	49.29	60.90
Ours	TRUTHDEBATE	**99.40**	**99.39**	**97.49**	**96.68**	**80.08**	**77.14**	**54.95**	**66.69**	**87.11**	**92.26**	**56.89**	66.82

Table 2: Performance comparison on the Washington Post and BBC domains and the average over all five test domains. The best result in each column is highlighted in bold.

Setting	Method	Wash.				BBC				AVG			
		ACC	F1	MaF1	mAP	ACC	F1	MaF1	mAP	ACC	F1	MaF1	mAP
Zero-shot	DeepSeek-VL2-27B	21.20	0.39	5.80	38.93	32.72	0.20	8.21	31.26	34.34	1.04	8.24	30.53
	Qwen2-VL-72B-Instruct	28.92	18.13	10.18	41.71	36.89	12.67	9.64	33.36	39.94	17.30	11.45	33.34
	Qwen3.6-27B	48.09	52.02	13.56	42.22	49.59	44.26	14.85	34.73	54.63	50.10	15.66	34.71
	Qwen3-VL-32B-Instruct	50.74	58.53	14.29	43.24	51.30	52.81	16.79	35.47	56.46	58.49	16.19	35.20
	Qwen3.5-122B-A10B	46.77	52.14	11.45	39.91	46.99	44.96	12.83	31.69	52.57	50.33	12.61	31.14
Supervised Models	ViLT	78.96	88.02	30.59	54.90	73.16	82.05	32.17	50.35	77.38	84.54	41.51	56.21
	HAMMER	79.81	88.53	32.22	66.19	76.09	83.78	36.05	58.36	78.51	85.52	41.67	66.24
	HAMMER++	80.28	88.67	30.98	66.21	76.71	83.68	34.41	58.83	79.53	85.65	42.14	67.48
	FKA-Owl	82.85	89.32	39.74	61.80	76.91	79.52	44.20	63.80	82.95	85.07	47.96	67.12
	AMD	82.85	88.35	46.72	62.59	82.65	86.26	39.44	50.44	83.69	86.10	50.37	61.41
Multi-Agent Baselines	Chain	85.35	90.85	53.64	66.97	84.08	87.34	54.62	65.22	83.97	86.72	58.62	67.97
	Star	85.42	90.94	55.10	65.52	83.61	87.43	52.13	61.32	81.93	85.63	56.95	64.55
	Tree	84.57	90.57	53.89	64.68	82.72	87.09	52.01	63.19	79.85	84.43	53.86	63.47
	Graph	84.33	90.48	50.80	63.14	81.15	85.71	49.81	63.16	78.82	83.59	51.47	63.11
	DyLAN	86.52	91.70	55.45	64.09	86.27	89.31	55.77	64.21	84.66	87.49	58.46	65.29
Ours	TRUTHDEBATE	**86.91**	**91.79**	**59.91**	**70.21**	**87.36**	**89.76**	**57.40**	**68.64**	**88.17**	**90.07**	**65.33**	**73.81**

recognition across the six fine-grained classes. We further report mAP over the three manipulation types: face swap, face emotion editing, and text fabrication. AVG denotes the unweighted mean over the five test domains. All predictive metrics are reported as percentages.

4.2 COMPARISON WITH BASELINES

Tables 1 and 2 compare TRUTHDEBATE with zero-shot models, NYT-supervised detectors, and multi-agent baselines across five news domains. Overall, fine-grained cross-domain recognition is more challenging than binary real/fake detection.

Zero-shot and supervised baselines. Among the zero-shot methods, Qwen3-VL-32B-Instruct achieves the best average performance, with ACC, F1, Macro-F1, and mAP scores of 56.46, 58.49, 16.19, and 35.20, respectively. Nevertheless, all zero-shot models obtain relatively low Macro-F1

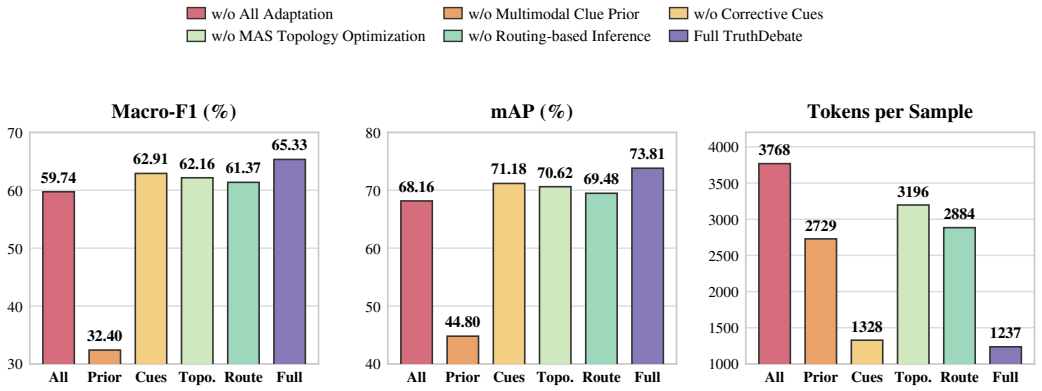

Figure 2: Component analysis of TRUTHDEBATE averaged over the five MDSM test domains in terms of Macro-F1, mAP, and token consumption per sample.

scores, indicating that direct prompting is insufficient for reliably recognizing fine-grained manipulation types across domains.

The NYT-supervised detectors substantially outperform the zero-shot models. AMD achieves the highest average ACC, F1, and Macro-F1 scores of 83.69, 86.10, and 50.37, respectively, while HAMMER++ obtains the highest average mAP of 67.48. However, their Macro-F1 scores decrease considerably on unseen domains, showing the difficulty of transferring fine-grained manipulation knowledge across news sources.

Comparison with multi-agent baselines. Among the topology-based baselines, Chain achieves the best average performance, with ACC, F1, Macro-F1, and mAP scores of 83.97, 86.72, 58.62, and 67.97, respectively. The differences among Chain, Star, Tree, and Graph show that multi-agent performance remains sensitive to the communication structure.

DyLAN dynamically selects participating agents and achieves average ACC, F1, Macro-F1, and mAP scores of 84.66, 87.49, 58.46, and 65.29, respectively. It obtains higher ACC and F1 than the topology-based baselines but remains below Chain in Macro-F1 and mAP. This result suggests that dynamic agent selection improves binary detection, while effective evidence integration remains important for fine-grained recognition.

TRUTHDEBATE achieves the best average performance across all four metrics, with ACC, F1, Macro-F1, and mAP scores of 88.17, 90.07, 65.33, and 73.81, respectively. Compared with the strongest supervised detector for each metric, TRUTHDEBATE improves ACC, F1, Macro-F1, and mAP by 4.48, 3.97, 14.96, and 6.33 percentage points. Compared with the strongest multi-agent baseline for each metric, the corresponding improvements are 3.51, 2.58, 6.71, and 5.84 percentage points.

The improvements remain consistent under domain shift. On Guardian, USA Today, Washington Post, and BBC, TRUTHDEBATE achieves Macro-F1 scores of 54.95, 56.89, 59.91, and 57.40, respectively, outperforming all multi-agent baselines. These results demonstrate the effectiveness of multimodal clue extraction and efficiency-aware multi-agent collaborative reasoning for both binary detection and fine-grained cross-domain recognition.

4.3 COMPONENT ANALYSIS

We conduct controlled ablations to evaluate the contributions of the multimodal clue prior, corrective cues, MAS topology optimization, and routing-based inference in TRUTHDEBATE. All variants use the same data splits, image–text inputs, visual and textual clues, underlying language model, and decoding configuration as the complete framework, except for the component explicitly removed in each ablation.

Figure 2 reports Macro-F1, mAP, and average token consumption over the five test domains. We compare six configurations. *Full* TRUTHDEBATE uses the complete framework. *w/o All Adaptation* jointly disables corrective cues, MAS topology optimization, and routing-based inference, and applies a fixed agent structure to every sample. *w/o Multimodal Clue Prior* removes the three-dimensional prior produced by the multimodal clue extraction model. Without the prior scores, direct prediction is unavailable and every sample is sent to full cooperative reasoning. The remaining variants separately remove corrective cues, MAS topology optimization, or routing-based inference.

Jointly removing all adaptation decreases Macro-F1 and mAP from 65.33 and 73.81 to 59.74 and 68.16, corresponding to reductions of 5.59 and 5.65 percentage points, respectively. Average token consumption increases from 1237.25 to 3768.20 tokens per sample, approximately $3.05\times$ the cost of the complete framework. These results show that the adaptive components jointly improve fine-grained recognition and inference efficiency.

Removing the multimodal clue prior causes the largest performance degradation. Macro-F1 decreases from 65.33 to 32.40, while mAP decreases from 73.81 to 44.80. Token consumption increases to 2728.87 per sample, approximately $2.21\times$ the cost of the complete framework. Without this prior, direct prediction is unavailable and the MAS lacks the prior knowledge needed to guide cooperative reasoning. This ablation therefore demonstrates the contribution of multimodal clue extraction to both detection performance and routing-based inference.

Without corrective cues, Macro-F1 and mAP decrease to 62.91 and 71.18, respectively, while token consumption increases moderately to 1328.40. This result shows that role-specific cues derived from training-time prediction errors improve subsequent evidence assessment with little additional online cost.

Removing MAS topology optimization decreases Macro-F1 and mAP to 62.16 and 70.62, respectively, while increasing token consumption to 3196.50, approximately $2.58\times$ that of the complete framework. Invoking all candidate analysts introduces redundant or less reliable evidence and increases both the number of language-model calls and the length of downstream prompts. MAS topology optimization improves performance and efficiency by retaining the most reliable text-related analysts.

Disabling routing-based inference decreases Macro-F1 and mAP to 61.37 and 69.48, respectively. It also increases average token consumption to 2883.60 tokens per sample, approximately $2.33\times$ the cost of the complete framework, because every sample is processed by full cooperative reasoning. These results show that applying the complete multi-agent procedure to samples already resolved confidently by the multimodal clue prior increases computational cost and may introduce unnecessary evidence.

Overall, TRUTHDEBATE achieves the highest Macro-F1 and mAP while consuming the fewest language-model tokens among all evaluated configurations. The results demonstrate that the multimodal clue prior, corrective cues, MAS topology optimization, and routing-based inference jointly improve fine-grained manipulation recognition and inference efficiency.

5 CONCLUSION

In this work, we presented TRUTHDEBATE, an adaptive multi-agent framework for deepfake detection. TRUTHDEBATE addresses the limited perspective of single-model analysis and the high cost of conventional multi-agent systems through two stages. Multimodal Clue Extraction captures visual and textual clues of deceptive content and produces a three-dimensional prior over face swap, face emotion editing, and text fabrication. Guided by this prior, Efficiency-aware Multi-Agent Collaborative Reasoning coordinates multimodal clue analysts, evidence integrators, an evidence auditor, and a decision-maker. It further combines training-time feedback-guided MAS optimization with routing-based inference to reduce unnecessary agent calls. Experiments on the MDSM dataset across five news domains show that TRUTHDEBATE outperforms supervised multimodal detectors and multi-agent baselines while reducing language-model token consumption. These results demonstrate the value of combining multimodal clue priors with adaptive cooperative reasoning for accurate and efficient deepfake detection.

REFERENCES

Biwei Cao, Qihang Wu, Jiuxin Cao, Bo Liu, and Jie 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.

Justin Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. MAGDi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 7220–7235. PMLR, 2024.

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

Hao Guo, Zihan Ma, Zhi Zeng, Minnan Luo, Weixin Zeng, Jiuyang Tang, and Xiang 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, 2025a.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–116. IEEE, 2025b.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Jen-Tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael Lyu, and Maarten Sap. On the resilience of LLM-based multi-agent collaboration with faulty agents. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 26202–26226. PMLR, 2025a.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841. IEEE, 2025b.

Lars Henning Klein, Nearchos Potamitis, Roland Aydin, Robert West, Caglar Gulcehre, and Akhil Arora. Fleet of agents: Coordinated problem solving with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 30986–31019. PMLR, 2025.

Haoran Li, Ziyi Su, Yun Xue, Zhiliang Tian, Yiping Song, and Minlie Huang. Advancing collaborative debates with role differentiation through multi-agent reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22655–22666, 2025a.

Yiheng Li, Yang Yang, Zichang Tan, Huan Liu, Weihua Chen, Xu Zhou, and Zhen Lei. Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9243–9252. IEEE, 2025b.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng 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.

Weihai Lu, Yu Tong, and Zhiqiu Ye. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 559–567, 2025.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang 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.

Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 6149–6157, 2020.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13052–13062. IEEE, 2024.

Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913. IEEE, 2023.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5556–5574, 2024.

Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. Learning to decode collaboratively with multiple language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12974–12990, 2024.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993, 2024.

Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, and Ping Luo. Forensics-bench: A comprehensive forgery detection benchmark suite for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4233–4245. IEEE, 2025.

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 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849–857, 2018.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 257–279, 2024.

Lianwei Wu, Pusheng Liu, and Yanning Zhang. See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 13736–13744, 2023.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25796–25804, 2025.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17591–17599, 2024a.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14544–14607, 2024b.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16777–16785, 2024c.

Yuchen Zhang, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. The coherence trap: When mllm-crafted narratives exploit manipulated visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8760–8769, 2026.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024d.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8580–8622, 2025a.

Ye Zhu, Yunan Wang, and Zitong 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, 2025b.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 62743–62767. PMLR, 2024.

A Use of Generative AI Tools

Generative AI tools were limited to minor editing tasks during manuscript preparation, such as refining sentence fluency and correcting typographical errors. They were not used to generate new content, analyses, or experimental results.

B Notation and Problem Formulation

For a news sample i, let $\mathcal{X}^{(i)} = (I^{(i)}, T^{(i)})$ denote its image–text pair, where $I^{(i)}$ and $T^{(i)}$ are the news image and associated text, respectively. The task query Q is shared across samples and omitted for simplicity.

We consider three manipulation dimensions:

$$\mathcal{D} = \{\mathrm{FS}, \mathrm{FE}, \mathrm{TF}\}, \tag{10}$$

corresponding to face swap, face emotion editing, and text fabrication. The dimension-wise target is

$$\mathbf{y}^{(i)} = \left[y_{FS}^{(i)}, y_{FE}^{(i)}, y_{TF}^{(i)}\right] \in \{0,1\}^3. \tag{11}$$

The Multimodal Clue Extraction model maps the image–text pair to a three-dimensional clue prior:

$$\mathbf{p}^{(i)} = \left[p_{FS}^{(i)}, p_{FE}^{(i)}, p_{TF}^{(i)}\right] \in [0,1]^3, \tag{12}$$

where $p_d^{(i)}$ is the predicted score for dimension $d \in \mathcal{D}$.

For samples requiring cooperative reasoning, each multimodal clue analyst or evidence integrator produces a dimension-wise report

$$o_{u,d}^{(i)} = \left(s_{u,d}^{(i)}, q_{u,d}^{(i)}, e_{u,d}^{(i)}\right), \tag{13}$$

where u denotes a reasoning role, $s_{u,d}^{(i)} \in \{\mathrm{supported}, \mathrm{unsupported}, \mathrm{uncertain}\}$ is the local judgment, $q_{u,d}^{(i)} \in [0,1]$ is the sample-specific confidence, and $e_{u,d}^{(i)}$ is a concise evidence explanation.

Table 3: Principal notation used in TRUTHDEBATE.

Notation	Definition	Notation	Definition
$\mathcal{X}^{(i)}$	The i-th image–text news sample	Q	Task query shared across samples
$\mathcal{D}$	Set of manipulation dimensions	$\mathbf{y}^{(i)}$	Ground-truth dimension-wise labels
$\mathbf{p}^{(i)}$	Three-dimensional multimodal clue prior	$p_d^{(i)}$	Prior score for dimension d
$\mathcal{D}_u$	Manipulation dimensions assigned to role u	$\mathcal{A}_{TF}$	Candidate text-related analyst set
$\mathcal{A}'_{TF}$	Selected top-K text-related analysts	$\mathcal{S}_t$	Active multimodal clue analyst set
$\mathcal{H}$	Set of evidence integrators	u	Generic reasoning role
$o_{u,d}^{(i)}$	Role u's report on dimension d	$s_{u,d}^{(i)}$	Categorical local judgment
$q_{u,d}^{(i)}$	Sample-specific report confidence	$e_{u,d}^{(i)}$	Concise evidence explanation
$c_{u,d}$	Number of correct feedback judgments	$n_{u,d}$	Number of valid feedback judgments
$r_{u,d}$	Feedback-derived role reliability	$\alpha_h^{(i)}$	Audit adjustment for integrator h
$w_{h,d}^{(i)}$	Effective integrator contribution weight	θ_d	Dimension-wise decision threshold
$\gamma_d^{(i)}$	Threshold-aware prior confidence	τ	Routing threshold
$g^{(i)}$	Indicator of cooperative reasoning	$\kappa_u^{(i)}$	Token cost of language-model role u
$C_{\text{full}}^{(i)}$	Cost of full cooperative reasoning	$C_{\text{TD}}^{(i)}$	Online token cost of TRUTHDEBATE

Let $\mathcal{A}_{TF}$ denote the candidate set of text-related multimodal clue analysts. The selected analyst set is denoted by $\mathcal{A}'_{TF} \subseteq \mathcal{A}_{TF}$, while $\mathcal{H}$ denotes the set of evidence integrators. For an integrator $h \in \mathcal{H}$, its contribution to the decision-maker on dimension d is weighted by

$$w_{h,d}^{(i)} = r_{h,d} \cdot q_{h,d}^{(i)} \cdot \alpha_h^{(i)}, \tag{14}$$

where $r_{h,d}$ is feedback-derived role reliability and $\alpha_h^{(i)} \in [0,1]$ is the adjustment assigned by the evidence auditor.

Table 3 summarizes the principal notation used in TRUTHDEBATE.

C Additional Method Details

C.1 Clue-Embedding Regularization

We regularize the learnable multimodal clue embeddings to encourage diverse clue representations and to prevent the feature energy of an embedding from concentrating on only a few dimensions. Let $\mathbf{E}_c^{(b)} \in \mathbb{R}^{N_c \times d}$ denote the contextualized clue-embedding matrix for sample b in a mini-batch of size B.

We first construct its token-wise Gram matrix:

$$\mathbf{G}^{(b)} = \mathbf{E}_c^{(b)} \left(\mathbf{E}_c^{(b)}\right)^\top. \tag{15}$$

The orthogonality term penalizes correlations between different clue embeddings:

$$\mathcal{L}_{\text{orth}} = \frac{1}{B} \sum_{b=1}^{B} \left\| \mathbf{G}^{(b)} - \text{Diag}\left(\text{diag}(\mathbf{G}^{(b)})\right) \right\|_F^2. \tag{16}$$

For the j-th clue embedding of sample b, its normalized feature energy at dimension ℓ is

$$v_{b,j,\ell} = \frac{e_{b,j,\ell}^2}{\sum_{m=1}^{d} e_{b,j,m}^2 + \epsilon}, \tag{17}$$

where $e_{b,j,\ell}$ is the ℓ-th feature and ϵ is a small constant. We encourage this distribution to remain close to uniform:

$$\mathcal{L}_{\text{mag}} = \frac{1}{BN_c} \sum_{b=1}^{B} \sum_{j=1}^{N_c} \sum_{\ell=1}^{d} v_{b,j,\ell} \log \left(\frac{v_{b,j,\ell} + \epsilon}{1/d} \right). \tag{18}$$

The clue-embedding regularization term in Equation 2 is

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{orth}} + \eta \mathcal{L}_{\text{mag}}, \tag{19}$$

where η controls the contribution of the feature-energy term.

C.2 Beta-smoothed Reliability Estimation

For each reasoning role u and manipulation dimension $d \in \mathcal{D}_u$, TruthDebate maintains a feedback-derived reliability estimate. Let $n_{u,d}$ denote the number of valid definite judgments made by role u on dimension d during training-time feedback, and let $c_{u,d}$ denote the number of judgments consistent with the ground-truth target. We use the posterior mean under a symmetric $\text{Beta}(2, 2)$ prior:

$$r_{u,d} = \frac{c_{u,d} + 2}{n_{u,d} + 4}. \tag{20}$$

For a training sample i, feedback is applied only to incorrectly predicted dimensions:

$$\mathcal{D}_{\text{err}}^{(i)} = \left\{ d \in \mathcal{D} \;\middle|\; \hat{y}_d^{(i)} \neq y_d^{(i)} \right\}. \tag{21}$$

For a definite role judgment, the corresponding binary prediction is

$$\widetilde{y}_{u,d}^{(i)} = \mathbb{I}\left[s_{u,d}^{(i)} = \text{supported} \right]. \tag{22}$$

For each $d \in \mathcal{D}_{\text{err}}^{(i)} \cap \mathcal{D}_u$, we update

$$\begin{aligned} n_{u,d} &\leftarrow n_{u,d} + 1, \\ c_{u,d} &\leftarrow c_{u,d} + \mathbb{I}\left[\widetilde{y}_{u,d}^{(i)} = y_d^{(i)} \right]. \end{aligned} \tag{23}$$

Reports marked as uncertain are excluded. The same update is applied to multimodal clue analysts and evidence integrators, while their statistics are maintained separately. Training feedback also updates role-level corrective cues for the involved analysts. All reliability estimates and corrective cues are frozen during test-time inference.

C.3 Top-K Analyst Selection

Let $\mathcal{A}_{TF}$ denote the candidate set of text-related multimodal clue analysts. During each training epoch, all candidates are evaluated, and the K analysts with the highest reliability on the text-fabrication dimension are retained:

$$\mathcal{A}'_{TF} = \text{TopK}_{a \in \mathcal{A}_{TF}} \left(r_{a,TF}; K \right). \tag{24}$$

The operator returns the identities of the selected analysts rather than their reliability values.

The selected text-related analysts are combined with the two always-active facial analysts:

$$\mathcal{S}_t = \{ a_{FS}, a_{FE} \} \cup \mathcal{A}'_{TF}. \tag{25}$$

The final analyst set is frozen after training and shared across all test samples. This selection does not explicitly optimize pairwise redundancy; limiting the active analyst set nevertheless reduces potentially overlapping reports, language-model calls, and the length of downstream integration prompts.

C.4 Cost-aware Routing Analysis

Let θ_d denote the decision threshold for manipulation dimension d. The clue-prior prediction is

$$\hat{y}^{(i)}_{d,\text{prior}} = \mathbb{I}\left[p^{(i)}_d \geq \theta_d\right]. \tag{26}$$

Because the decision thresholds may differ across dimensions, prior confidence is measured relative to the corresponding threshold:

$$\gamma^{(i)}_d = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\dfrac{p^{(i)}_d - \theta_d}{1-\theta_d}, & p^{(i)}_d \geq \theta_d, \\ \dfrac{1}{2} + \dfrac{1}{2}\dfrac{\theta_d - p^{(i)}_d}{\theta_d}, & p^{(i)}_d < \theta_d. \end{cases} \tag{27}$$

Thus, $\gamma^{(i)}_d = 0.5$ at the decision boundary and increases as the prior score moves farther from that boundary.

Given routing threshold τ, the routing indicator is

$$g^{(i)} = \mathbb{I}\left[\min_{d\in\mathcal{D}} \gamma^{(i)}_d < \tau\right]. \tag{28}$$

When $g^{(i)} = 0$, the system directly uses the three-dimensional clue prior. When $g^{(i)} = 1$, the selected analysts, evidence integrators, evidence auditor, and decision-maker are invoked.

Let $\kappa^{(i)}_u$ denote the total input and output token cost of language-model role u for sample i. The cost of full cooperative reasoning is

$$C^{(i)}_{\text{full}} = \sum_{a\in\mathcal{S}_t} \kappa^{(i)}_a + \sum_{h\in\mathcal{H}} \kappa^{(i)}_h + \kappa^{(i)}_{\text{audit}} + \kappa^{(i)}_{\text{dec}}. \tag{29}$$

The online language-model token cost of TruthDebate is therefore

$$C^{(i)}_{\text{TD}} = g^{(i)} C^{(i)}_{\text{full}}. \tag{30}$$

The Multimodal Clue Extraction model and precomputed evidence do not invoke the language model during inference and are excluded from this token count.

C.5 Algorithms

Algorithm 1 summarizes the complete training and test-time workflow of TruthDebate.

D Additional Experimental Details

D.1 Implementation Details

The Multimodal Clue Extraction model uses 32 learnable multimodal clue embeddings. It is optimized with AdamW for 12 epochs using a batch size of 8, weight decay of 1×10^{-4}, and learning rates of 1×10^{-6} and 1×10^{-4}. Input images are resized to 768×768.

For multi-agent inference, five samples are processed in each batch. The temperature is set to 0.05 for multimodal clue analysts and 0.1 for evidence integration and decision-making. We set the routing threshold to $\tau = 0.90$ and retain $K = 2$ text-related analysts. The routing threshold, dimension-wise decision thresholds, and other model-selection choices are determined on the NYT validation split and fixed for all five test domains.

D.2 Baselines

We compare TruthDebate with zero-shot multimodal models, supervised multimodal detectors, and multi-agent baselines.

Algorithm 1 Training and inference procedure of TRUTHDEBATE

Require: Training set $\mathcal{S}_{\mathrm{tr}}$, test sample $(I^{(i)}, T^{(i)})$, Multimodal Clue Extraction model f_{clue}, candidate text-related analyst set $\mathcal{A}_{TF}$, evidence integrators $\mathcal{H}$, training epochs N_{ep}, selection budget K, decision thresholds $\{\theta_d\}_{d\in\mathcal{D}}$, and routing threshold τ

Ensure: Final dimension-wise prediction $\hat{\mathbf{y}}^{(i)}$

Training-time feedback adaptation and MAS optimization:
Initialize $c_{u,d} \leftarrow 0$, $n_{u,d} \leftarrow 0$, and $r_{u,d} \leftarrow 0.5$ for all applicable (u, d)
Initialize the training analyst set $\mathcal{S}_{\text{tr-agent}} \leftarrow \{a_{FS}, a_{FE}\} \cup \mathcal{A}_{TF}$
Initialize $\mathcal{S}_t \leftarrow \mathcal{S}_{\text{tr-agent}}$
for $e = 1$ to N_{ep} **do**
    for each training sample $(I^{(j)}, T^{(j)}, \mathbf{y}^{(j)}) \in \mathcal{S}_{\mathrm{tr}}$ **do**
        Obtain clue prior $\mathbf{p}^{(j)} \leftarrow f_{\mathrm{clue}}(I^{(j)}, T^{(j)}, Q)$
        Construct a deterministic clue-prior report from $\mathbf{p}^{(j)}$
        Collect reports from all candidate analysts in $\mathcal{S}_{\text{tr-agent}}$
        Integrate the analyst and clue-prior reports through $\mathcal{H}$ and audit the resulting evidence
        Obtain final prediction $\hat{\mathbf{y}}^{(j)}$
        $\mathcal{D}_{\mathrm{err}}^{(j)} \leftarrow \{d \in \mathcal{D} \mid \hat{y}_d^{(j)} \neq y_d^{(j)}\}$
        for each $d \in \mathcal{D}_{\mathrm{err}}^{(j)}$ **do**
            for each role u with a definite report on d **do**
                Update $n_{u,d}$, $c_{u,d}$, and $r_{u,d}$ using Equation 23
            end for
            Update corrective cues for involved analysts
        end for
    end for
    Rank all $a \in \mathcal{A}_{TF}$ by $r_{a,TF}$
    $\mathcal{A}'_{TF} \leftarrow \mathrm{TopK}_{a\in\mathcal{A}_{TF}}(r_{a,TF}; K)$
    $\mathcal{S}_t \leftarrow \{a_{FS}, a_{FE}\} \cup \mathcal{A}'_{TF}$
end for
Freeze $\{r_{u,d}\}$, corrective cues, and $\mathcal{S}_t$

Routing-based test-time inference:
Obtain clue prior $\mathbf{p}^{(i)} \leftarrow f_{\mathrm{clue}}(I^{(i)}, T^{(i)}, Q)$
for each $d \in \mathcal{D}$ **do**
    Compute $\hat{y}_{d,\mathrm{prior}}^{(i)}$ using Equation 26
    Compute $\gamma_d^{(i)}$ using Equation 27
end for
if $\min_{d\in\mathcal{D}} \gamma_d^{(i)} \geq \tau$ **then**
    return $\{\hat{y}_{d,\mathrm{prior}}^{(i)}\}_{d\in\mathcal{D}}$
end if
Construct a deterministic clue-prior report from $\mathbf{p}^{(i)}$
Query each selected analyst $a \in \mathcal{S}_t$ using its assigned precomputed evidence and frozen corrective cue
Integrate the analyst and clue-prior reports through $\mathcal{H}$
Audit the reports and obtain $\alpha_h^{(i)}$ for each integrator h
for each integrator $h \in \mathcal{H}$ and $d \in \mathcal{D}_h$ **do**
    $w_{h,d}^{(i)} \leftarrow r_{h,d} q_{h,d}^{(i)} \alpha_h^{(i)}$
end for
Fuse the weighted integrator reports through the decision-maker
Obtain and return $\hat{\mathbf{y}}^{(i)}$

Zero-shot multimodal models. We evaluate DeepSeek-VL2-27B, Qwen2-VL-72B-Instruct, Qwen3.6-27B, Qwen3-VL-32B-Instruct, and Qwen3.5-122B-A10B without task-specific fine-tuning. Each model receives the news image and associated text through a unified instruction prompt

Table 4: Implementation hyperparameters.

Hyperparameter	Value
Number of learnable clue embeddings	32
Optimizer	AdamW
Learning rate	1×10^{-6} / 1×10^{-4}
Weight decay	1×10^{-4}
Training batch size	8
Training epochs	12
Image resolution	768×768
Agent batch size	5
Analyst temperature	0.05
Integration and decision temperature	0.1
Routing threshold τ	0.90
Number of retained analysts K	2

and directly predicts one of the six fine-grained classes. The prompt specifies the class definitions and requires a structured prediction, but does not provide the three-dimensional clue prior, precomputed analyst evidence, or training-time feedback information.

Supervised multimodal detectors. We compare with ViLT, HAMMER, HAMMER++, FKA-Owl, and AMD. ViLT is a vision–language transformer baseline that jointly encodes image and text inputs. HAMMER and HAMMER++ are manipulation-aware multimodal detectors designed for media-manipulation classification and grounding. FKA-Owl augments a large vision–language model with external knowledge for multimodal fake news detection, while AMD is a supervised multimodal detector with fine-grained manipulation prediction. All supervised methods are trained on the NYT training split. We select checkpoints using the NYT validation split and evaluate them on the same in-domain and cross-domain test sets as TRUTHDEBATE. For each baseline, we retain its original input representation and prediction protocol whenever applicable.

Multi-agent baselines. We compare with Chain, Star, Tree, Graph, and DyLAN. To control for the underlying model and evidence source, all multi-agent baselines use the same Qwen3.6-27B language model, three-dimensional clue prior, precomputed multimodal evidence, candidate analyst roles, data splits, and decoding configuration as TRUTHDEBATE.

Chain sequentially passes reports from one analyst to the next. Star sends all analyst reports to a central decision agent. Tree performs hierarchical aggregation from bottom-level analysts to intermediate nodes and then to a root decision agent. Graph enables message passing among connected agents under a predefined graph structure. These four baselines use fixed communication topologies. DyLAN dynamically selects participating agents for each sample following its original agent-selection procedure.

For the fixed-topology baselines, we vary only the communication structure and do not apply the proposed feedback-guided correction, Top-K analyst selection, or uncertainty-aware routing. This setup isolates the effect of conventional multi-agent communication topologies. DyLAN provides a dynamic-selection comparison, whereas TRUTHDEBATE additionally combines feedback-derived role reliability, analyst selection, evidence auditing, and confidence-based routing.

D.3 EVALUATION METRICS

Let $b^{(i)} \in \{0, 1\}$ and $\hat{b}^{(i)} \in \{0, 1\}$ denote the binary real/fake target and prediction of sample i, respectively, where all manipulated samples are treated as the positive class. Binary accuracy is

$$\mathrm{ACC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{b}^{(i)} = b^{(i)}\right]. \tag{31}$$

Let TP, FP, and FN denote the true-positive, false-positive, and false-negative counts for the manipulated class. Binary F1 is

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{32}$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{33}$$

Undefined precision, recall, or F1 values caused by a zero denominator are set to zero.

For six-class fine-grained recognition, Macro-F1 is the unweighted mean of class-wise one-vs-rest F1 scores:

$$\text{MaF1} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \text{F1}_y, \qquad |\mathcal{Y}| = 6, \tag{34}$$

where $\mathcal{Y}$ contains the authentic class and five manipulation classes.

For each manipulation dimension $d \in \mathcal{D}$, let $\widehat{p}_d^{(i)} \in [0, 1]$ be the final predicted score and $y_d^{(i)} \in \{0, 1\}$ be the ground-truth indicator. After ranking samples by $\widehat{p}_d^{(i)}$ in descending order, the average precision is

$$\text{AP}_d = \frac{1}{N_d^+} \sum_{k=1}^{N} \text{Precision}_d(k) \cdot \mathbb{I}\left[y_d^{(\pi_d(k))} = 1\right], \tag{35}$$

where $\pi_d(k)$ is the index ranked at position k, $\text{Precision}_d(k)$ is the precision among the top-k samples, and N_d^+ is the number of positive samples for dimension d. The mean average precision is

$$\text{mAP} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \text{AP}_d. \tag{36}$$

For each predictive metric M, AVG denotes the unweighted mean over the five test domains:

$$\text{AVG}(M) = \frac{1}{5} \sum_{j=1}^{5} M_j. \tag{37}$$

All predictive metrics are reported as percentages. Tokens denotes the average total number of language-model input and output tokens consumed per test sample.

E Additional Results and Analysis

E.1 Additional Binary Detection Analysis

The main-paper component analysis focuses on Macro-F1, mAP, and token consumption. Here, we additionally report average binary real/fake detection performance over the five MDSM test domains.

Figure 3 shows that the complete TruthDebate framework achieves the highest average ACC and F1, with 88.17 and 90.07, respectively. Removing all adaptive mechanisms reduces ACC and F1 to 84.92 and 87.02, indicating that feedback-guided correction, analyst selection, and uncertainty-aware routing jointly improve binary detection.

Removing the multimodal clue prior causes the largest degradation, reducing ACC to 55.20 and F1 to 49.10. This result shows that the clue prior provides essential guidance for distinguishing authentic and manipulated samples before cooperative reasoning.

The individual adaptive mechanisms also contribute consistently. Removing feedback-guided correction, Top-K analyst selection, or uncertainty-aware routing reduces ACC/F1 to 86.71/88.42, 86.38/88.10, and 85.84/87.68, respectively. These results are consistent with the fine-grained evaluation: the multimodal clue prior and adaptive cooperative reasoning improve both binary and fine-grained deepfake detection.

E.2 Token Cost Analysis

Figure 4 compares the average online language-model token cost of TruthDebate and its ablated variants. The Multimodal Clue Extraction model and precomputed evidence do not invoke the language model during inference and are excluded from this measure.

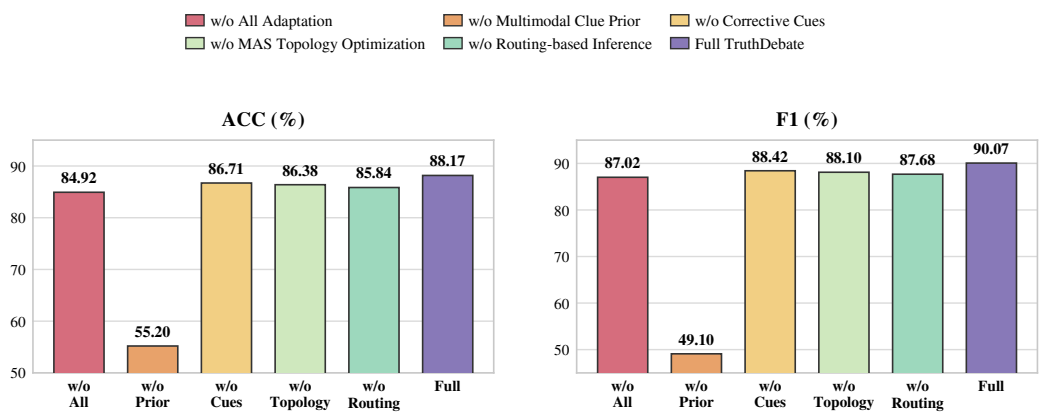

Figure 3: Binary detection performance of TRUTHDEBATE and its ablated variants, averaged over the five MDSM test domains.

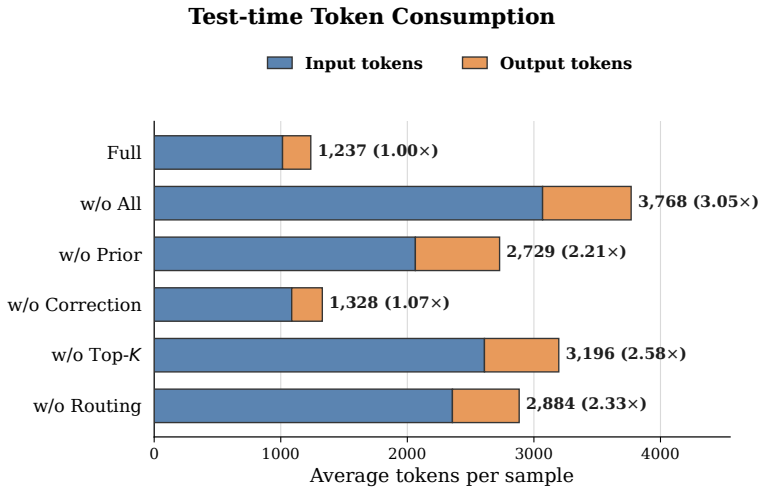

Figure 4: Average online language-model token consumption per sample. Each bar separates input and output tokens, and the total cost is reported at the right of the bar.

The complete framework consumes 1237.25 tokens per sample, including 1014.09 input tokens and 223.16 output tokens. Disabling all adaptive mechanisms increases the cost to 3768.20 tokens, approximately $3.05\times$ that of the complete framework. Thus, the combined adaptive mechanisms reduce online token consumption by approximately 67.2% relative to the static full-reasoning configuration.

Removing the multimodal clue prior increases the cost to 2728.87 tokens per sample. Without the prior, direct prediction is unavailable and all samples require cooperative reasoning. Removing feedback-guided correction results in a comparatively small increase to 1328.40 tokens, which is consistent with its primary role in improving evidence quality rather than eliminating agent calls.

Top-K analyst selection and uncertainty-aware routing provide the largest individual efficiency gains. Without analyst selection, the cost increases to 3196.50 tokens because all candidate analysts are invoked and additional reports are passed to downstream integrators. Without routing, the cost increases to 2883.60 tokens because every sample is processed through the full cooperative reasoning path. These results show that analyst selection reduces communication within the reasoning path, whereas routing allows high-confidence samples to bypass that path entirely.

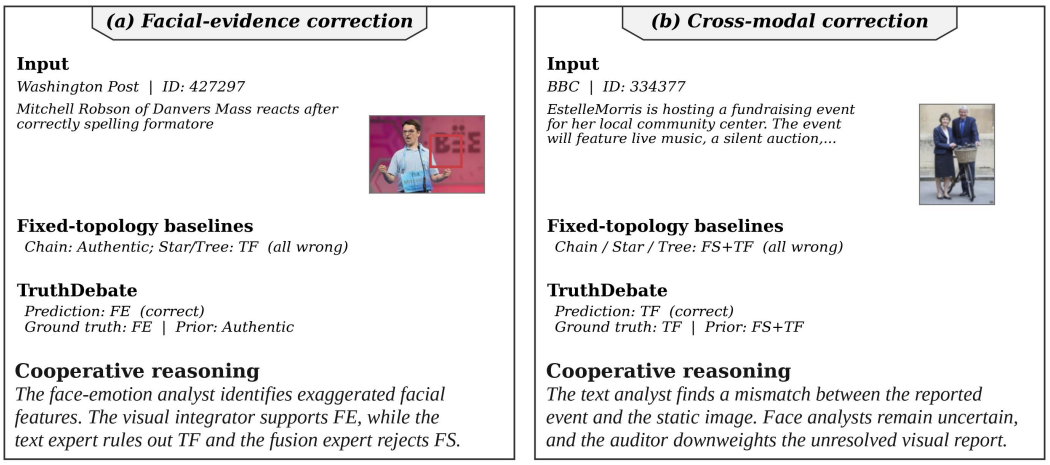

Figure 5: Qualitative comparison with fixed-topology multi-agent baselines. Chain, Star, and Tree make incorrect fine-grained predictions on both samples, whereas TRUTHDEBATE reaches the ground-truth result through clue-guided cooperative reasoning. Red boxes denote annotated manipulated regions.

E.3 QUALITATIVE COMPARISON WITH BASELINES

Figure 5 compares TRUTHDEBATE with three fixed-topology multi-agent baselines: Chain, Star, and Tree. The two samples are selected because all displayed baselines make incorrect fine-grained predictions, whereas TRUTHDEBATE correctly identifies the manipulation type.

In Figure 5(a), the multimodal clue prior predicts the authentic class, and the baselines predict either the authentic class or text fabrication. The face-emotion analyst, however, identifies exaggerated facial features. The visual integrator therefore supports face emotion editing, while the text expert rules out text fabrication and the fusion expert rejects face swapping. TRUTHDEBATE consequently produces the correct face-emotion-editing prediction.

In Figure 5(b), the prior and all three baselines predict face swapping together with text fabrication. The text analyst identifies a discrepancy between the reported live-music and silent-auction event and the static visual content. Meanwhile, the face-related analysts remain uncertain about facial manipulation. The evidence auditor downweights the unresolved visual report, and the final decision retains text fabrication only.

These cases illustrate that role-specific evidence analysis, uncertainty handling, and evidence auditing can correct misleading prior signals and fixed-topology reasoning errors.

F PROMPT TEMPLATES

Multimodal Clue Analyst Prompt

System Prompt. You are ⟨ROLE⟩, a multimodal clue analyst. Assess only ⟨DIMENSIONS⟩ using the assigned precomputed evidence. The evidence may be noisy or incomplete. Confidence represents the quality of the evidence for the current sample rather than the permanent ability of the analyst.
Do not infer manipulation dimensions outside your assigned scope. For every assigned dimension, output a state from {supported, unsupported, uncertain} and a confidence score between 0 and 1.
Provide a concise, evidence-grounded reason without introducing unsupported information.
When available, use the following corrective cue to avoid previously observed role-specific errors: ⟨ROLE-SPECIFIC CORRECTIVE CUE⟩. The corrective cue describes only a historical failure pattern and never contains the ground-truth label of the current sample.
Return only a JSON list in the same order as the input samples. Each item must contain the sample identifier, dimension-wise assessments, confidence scores, and a one-sentence reason.

User Input. A JSON list containing ⟨SAMPLE ID⟩ and ⟨PRECOMPUTED EVIDENCE⟩ for each sample.

Evidence Integrator Prompt

System Prompt. You are ⟨EXPERT ROLE⟩, an evidence integrator. Decide only the manipulation dimensions ⟨DIMENSIONS⟩. For every sample, evaluate the reports produced by your assigned child agents and estimate the confidence of your dimension-wise assessment.
Historical reliability describes the long-term quality of a child agent, whereas report confidence describes the strength of its evidence for the current sample. Consider these quantities separately. Do not treat duplicated reports or repeated references to the same underlying evidence as independent support.
For each assigned dimension, output a state from {supported, unsupported, uncertain} and a confidence score between 0 and 1. Provide a concise reason that identifies the evidence supporting the assessment and any unresolved conflict.
Return only a JSON list in the same order as the input samples. Each item must contain the sample identifier, dimension-wise assessments, confidence scores, and a one-sentence reason.

User Input. A JSON list containing ⟨SAMPLE ID⟩, ⟨CHILD-AGENT REPORTS⟩, and ⟨DIMENSION-WISE HISTORICAL RELIABILITIES⟩.

Evidence Auditor Prompt

System Prompt. You are an evidence auditor, not a final decision-maker. Examine the bottom-level and evidence-integrator reports for missing, weak, circular, duplicated, conflicting, or unsupported evidence. Treat multiple references to the same underlying evidence as a single source rather than independent support. Evaluate evidence completeness, source independence, inter-integrator consistency, and the agreement between reported confidence and evidence strength.
For each evidence integrator, return an adjustment score between 0 and 1. A lower score indicates duplicated, insufficient, or conflicting evidence, whereas a higher score indicates a complete and coherent evidence chain.
Return only a JSON list in the same order as the input samples. Each item must contain the sample identifier, integrator-specific adjustment scores, duplicated evidence sources, identified conflicts, and a one-sentence reason.

User Input. A JSON list containing ⟨SAMPLE ID⟩, ⟨BOTTOM-LEVEL REPORTS⟩, and ⟨EVIDENCE-INTEGRATOR REPORTS⟩.

Decision-maker Prompt

System Prompt. You are the decision-maker. Decide each manipulation dimension separately using the evidence-integrator reports, their effective weights, and the evidence-audit results.
Respect the audit adjustments. Give greater consideration to reports from historically reliable roles when their current evidence is strong and coherent. Do not count duplicated evidence multiple times. When evidence is insufficient or conflicting, return an uncertain assessment rather than introducing unsupported claims.
For each dimension, output a state from {supported, unsupported, uncertain} and a confidence score between 0 and 1. Provide a concise evidence-grounded reason.
Return only a JSON list in the same order as the input samples. Each item must contain the sample identifier, dimension-wise assessments, confidence scores, and a one-sentence reason.

User Input. A JSON list containing ⟨SAMPLE ID⟩, ⟨EVIDENCE-INTEGRATOR REPORTS⟩, ⟨EVIDENCE-AUDIT RESULTS⟩, and ⟨EFFECTIVE INTEGRATOR WEIGHTS⟩.