Hierarchical Progressive Alignment: A Cognitive-inspired Framework for Short-Video Fake News Detection

Anonymous ACL submission

Abstract

Fake News Detection is a very important but challenging task that has attracted the attention of both the field of natural language processing and multimedia computing. Most existing approaches to fake news detection on short-video platforms adopt static cross-modal fusion, directly combining visual content with auxiliary modalities such as text and audio for classification. While effective in some cases, this paradigm is sensitive to modality-specific noise and tends to overfit superficial cross-modal correlations, particularly on easy samples, which can undermine robustness. To address these issues, we introduce **HPA**, a diffusion-driven **Hierarchical Progressive Alignment** framework that performs adaptive computation. HPA begins with a lightweight authenticity predictor that produces an initial decision along with a confidence estimate, and selectively routes uncertain samples to subsequent stages. For these samples, a diffusion-based opinion evolution module iteratively denoises and reconstructs modality-specific semantic opinions, encouraging alignment in a shared latent space through a reconstruction objective. A fine-grained attribution module then refines the final prediction and provides cues associated with potential manipulation. Experimental results on **FakeSV** and **FakeTT** show that HPA achieves consistent improvements over strong baselines and generalizes well across datasets.

1 Introduction

Short-video platforms have become a major channel for the creation and spread of fake news (Bu et al., 2023; Sundar et al., 2021). Compared with purely textual misinformation, short-video fake news often mixes textual headlines and subtitles, visual content, and audio commentary or ambient sound (Hou et al., 2019; Papadopoulou et al., 2018; Qi et al., 2022; Bu et al., 2024), while manipulating only a subset of these modalities. This makes detection particularly challenging: a forged headline may be paired with authentic footage, or synthetic audio may be overlaid on real scenes, leading to subtle but harmful inconsistencies.

The proliferation of short-video fake news has evolved into a complex spectrum of deception (Niu et al., 2023), ranging from sensationally absurd content to highly sophisticated, localized tampering. As illustrated in Figure 1, deceptive cues vary drastically in perceptibility: a video claiming "UFOs landing" (Figure 1(a)) triggers immediate disbelief due to its violation of common sense, whereas a manipulated news clip about primary school students (Figure 1(b)) appears authentic in its visual footage but contains subtle textual tampering in the subtitles. Existing multimodal detection frameworks often adopt a "static fusion" paradigm (Qi et al., 2022; Wang et al., 2024, 2025a; Choi and Ko, 2021; Shang et al., 2021), compressing text, audio, and visual signals into a unified representation, this approach uses a uniform inference strategy for all cases. As a result, localized tampering cues (e.g., the edited text in Figure 1(b)) can be overwhelmed by authentic video/audio signals, causing failures on subtle forgeries(Zong et al., 2024).

We draw inspiration from Dual-System Theory (Kahneman, 2011), which posits two complementary modes of cognition: System 1 ("fast thinking") operates intuitively to recognize patterns rapidly, whereas System 2 ("slow thinking") engages in deliberate, analytical reasoning to resolve ambiguity. We propose the **Hierarchical Progressive Alignment (HPA)** framework to instantiate this cognitive pipeline. For the initial phase, HPA employs a detection mechanism analogous to Fast Thinking. For highly conspicuous cases like the "100% UFO" example in Figure 1(a), the model utilizes a coarse-grained Mixture-of-Experts (MoE)(Jacobs et al., 1991)to capture salient semantic conflicts and visual anomalies, enabling rapid high-confidence rejection without deep forensic analysis .

However, the core challenge lies in ambiguous

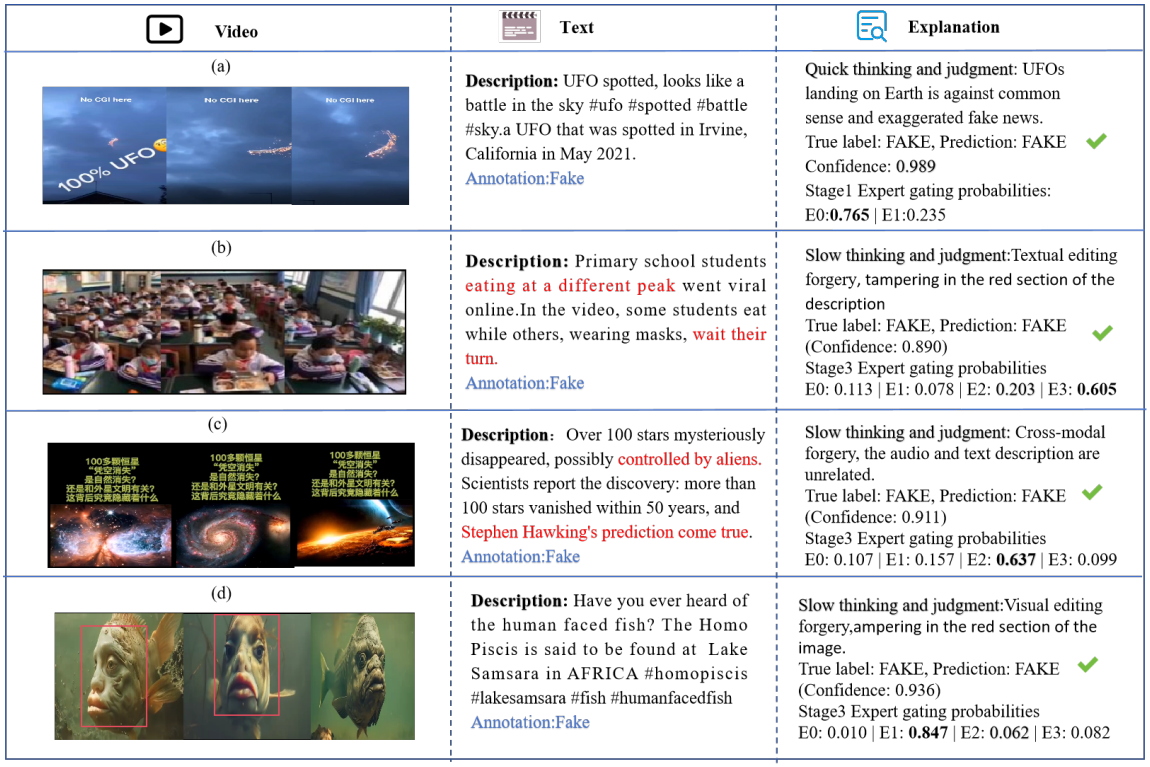


Figure 1: Examples of fake news in short videos illustrating the need for dual-system reasoning. (a) **Fast-thinking case:** An obvious fake due to absurd "UFO" claims defying common sense. (b–d) **Slow-thinking cases:** Subtle deceptions requiring detailed analysis, including (b) textual editing (red text), (c) cross-modal inconsistency (audio-text mismatch), and (d) visual tampering (red bounding boxes). Our HPA model handles these via cognitive-inspired stages corresponding to Fast and Slow thinking.

"Slow Thinking" scenarios (Figure 1(b–d)), where the deception is localized and conflicting evidence exists across modalities. To address this, we introduce **Stage 2: Opinion Evolution via Diffusion**. In cases like Figure 1(b), strictly static models often classify the video as "Real" because the visual and audio streams are genuinely authentic. To correct this, our diffusion module simulates the dynamic process of human hesitation and re-evaluation. By injecting noise into single-modal judgments to model cognitive uncertainty and then performing a cross-modal denoising process, the model propagates the deceptive signal from the tampered text (e.g., "eating at a different peak") to the other modalities. This "opinion evolution" forces the model to align the visual and audio representations with the textual inconsistency, effectively amplifying the localized fake signal so it is not drowned out by the authentic majority.

Finally, to mirror the expert analytical capability of System 2, HPA triggers **Stage 3: Manipulation MoE** for fine-grained diagnosis. Once a sample undergoes the slow thinking process, specific experts are activated to pinpoint the source of manipulation, as demonstrated in the contrasting cases of Figure 1. For the student dining video (Figure 1(b)), the *Textual Manipulation Expert* focuses on artifact tokens in the description to identify the semantic fabrication. Conversely, for the "human-faced fish" (Figure 1(d)), the *Visual Manipulation Expert* is gated to attend to the tampered regions (marked in red boxes), while the *Cross-modal Expert* handles cases like Figure 1(c), where the audio narration about "vanishing stars" ostensibly mismatches the background visuals. This stage ensures that the model provides not just a binary decision, but an interpretable attribution of *why* the content is fake.

In summary, our contributions are fourfold:

- **Cognitive Alignment:** We propose the HPA framework, the first to explicitly model Kahneman's dual-system theory in short-video fake news detection, handling both "Fast" intuition and "Slow" reasoning.
- **Opinion Evolution via Diffusion:** We in-

troduce a novel diffusion-based interaction mechanism for Stage 2, which solves the problem of dominant authentic signals by propagating localized tampering cues across modalities.

- **Interpretable Attribution:** We design a modality-specific Manipulation MoE that offers granular explanations for diverse forgery types (textual, visual, or cross-modal), significantly enhancing model transparency.

- **Balanced Effectiveness and Efficiency:** Extensive experiments on short-video fake news benchmarks demonstrate that HPA achieves competitive (often superior) detection performance while remaining lightweight and efficient, striking a strong balance between accuracy and practical deployment cost.

2 Related Work

2.1 Multimodal Fake News Detection and Dynamic Interaction

Early exploration of multimodal fake news detection focused on image-text pairs,(Wang et al., 2018) introduced event adversarial training to disentangle event-specific noise, while Zhou et al. (2020) quantified cross-modal similarity via cosine distance, treating all inconsistencies uniformly. These methods laid the foundation for consistency modeling but failed to handle nuanced conflicts. Models like SV-FEND (Qi et al., 2022) and others utilizing attention mechanisms (Wang et al., 2024, 2025a; Choi and Ko, 2021; Shang et al., 2021) typically concatenate visual, acoustic, and textual features into a unified representation. While effective for global inconsistencies, these methods often struggle with *localized* tampering, where authentic signals from unchanged modalities overwhelm the deceptive cues.

Opinion dynamics studies how individuals update their beliefs by partially incorporating others' opinions, and such interactions can converge to stable patterns such as consensus, polarization, or fragmentation (Dong et al., 2018) . Classic models, including DeGroot (DeGroot, 1974), the voter model(Ben-Naim et al., 1996), and continuous-opinion/discrete-action formulations (Martins, 2008, 2014), provide foundational fusion rules for describing opinion evolution.The diffusion model is a neural generative model based on the stochastic diffusion process. Recent developments in the diffusion model have primarily focused on generative tasks,such as image generation,natural language generation and audio generation Ho et al. (2020); Popov et al. (2021); Dhariwal and Nichol (2021)

To address the limitations of static fusion,Zong et al. (2024) have begun to model the dynamic interaction between modalities. Zong et al. (2024)pioneered the concept of Opinion Evolution, introducing a diffusion-based mechanism that injects noise into unimodal representations to simulate the hesitation and re-evaluation process of human judgment. By iteratively denoising, this approach effectively amplifies subtle conflicts between modalities. However, applying such a computationally expensive diffusion process to *all* input samples including obvious fakes that defy common sense is inefficient. Our HPA framework adopts this diffusion-based interaction specifically for the "Slow Thinking" stage, ensuring that complex reasoning is allocated only when necessary, while obvious cases are handled by a rapid rejection mechanism.

2.2 Adaptive Reasoning via MoE

The complexity of short-video misinformation requires models that can adaptively select reasoning pathways. Mixture-of-Experts (MoE)(Jacobs et al., 1991) architectures have gained traction for their ability to specialize in different data patterns.Shazeer et al. (2017); Jacobs et al. (1991); Fedus et al. (2022) Recent work on (Ying et al., 2022; Zhu et al., 2025) explores consistency across auxiliary views to refine unimodal features. This method uses contrastive learning to align diverse views, improving robustness to modality imbalance. However, it treats all views uniformly and does not prioritize cognitive effort allocation for ambiguous cases.Liu et al. (2025) proposed **MIMoE-FND**, a hierarchical MoE framework that routes samples based on "modality interaction types" (e.g., Agreed Alignment vs. Disagreed Misalignment). By analyzing the consistency between unimodal predictions, MIMoE dynamically activates specific experts to handle conflicting signals.

While MIMoE focuses on routing based on *signal consistency*, it treats the detection process as a single-step classification problem. Wang et al. (2025b) proposed **FakeSV-VLM**, a Vision-Language Model (VLM) augmented with a Progressive MoE (PMOE) adapter. PMOE decom-

poses reasoning into two stages: a *Detection MoE* for binary judgment and an *Attribution MoE* for identifying manipulation types (e.g., visual splicing). By leveraging VLM's semantic knowledge, FakeSV-VLM enhances interpretability but uses a static adapter structure, which cannot simulate dynamic cognitive processes like hesitation.In contrast, our work draws inspiration from the *Dual-System Theory* (Kahneman, 2011) to model the *cognitive hierarchy* of detection. Instead of routing solely based on conflict, our HPA framework differentiates between "Fast" intuition (for obvious absurdities) and "Slow" analytical reasoning (for subtle forgeries). furthermore, unlike MIMoE's abstract fusion experts, our Stage-3 Manipulation MoE is explicitly grounded in forensic categories (Textual, Visual, and Cross-modal Experts), providing not just a binary verdict but granulized explanations for *why* a video is fake.

3 Method

We propose a **Hierarchical Progressive Attribution (HPA)** framework for multimodal fake news detection on short video platforms (Fig. 2). HPA contains three staged predictors: (i) a lightweight **DetectionMoE** for coarse prediction and confidence-based routing, (ii) **Opinion Evolution via Diffusion** that evolves cross-modal "opinions" for hard samples, and (iii) **ManipulationMoE** for manipulation-aware reasoning with token-level attributions.

3.1 Problem Formulation

Given a dataset $\{(x_i, y_i)\}_{i=1}^{N}$, each sample x_i contains text x_i^t (title/ASR), audio x_i^a, and visual x_i^v (key frames/clips). The label $y_i \in \{0, 1\}$ indicates real or fake news. We learn a classifier:

$$\hat{y}_i = f_\theta(x_i), \quad x_i = \{x_i^t, x_i^a, x_i^v\}. \tag{1}$$

3.2 Multimodal Encoders

We embed each modality into a shared d-dimensional space:

$$f_m = \phi_m(x^m) \in R^d, \quad m \in \{t, a, v\}, \tag{2}$$

where ϕ_t, ϕ_a, ϕ_v are lightweight encoders with mean pooling for sequence inputs.

3.3 Stage I: DetectionMoE and Confidence Routing

We concatenate modal features:

$$z_{\text{det}} = [f_t; f_a; f_v] \in R^{3d}. \tag{3}$$

DetectionMoE contains K experts $\{E_k\}_{k=1}^{K}$ and a gating network:

$$\boldsymbol{\pi} = \text{softmax}(W_g z_{\text{det}} + \mathbf{b}_g) \in R^K. \tag{4}$$

The coarse prediction is computed by a weighted expert mixture:

$$\hat{\mathbf{y}}^{(1)} = \text{softmax}\left(W_c \sum_{k=1}^{K} \pi_k E_k(z_{\text{det}}) + \mathbf{b}_c \right). \tag{5}$$

We define sample confidence as $\text{conf} = \max(\hat{\mathbf{y}}^{(1)})$. Given a threshold τ, samples with $\text{conf} > \tau$ are treated as *easy* and directly output $\hat{\mathbf{y}}^{(1)}$; otherwise they are routed to Stage II–III.

3.4 Stage II: Opinion Evolution via Diffusion

For hard samples, we treat unimodal features as initial opinions:

$$x_0^t = f_t, \quad x_0^a = f_a, \quad x_0^v = f_v. \tag{6}$$

We first form a global multimodal opinion:

$$x_0^{\text{multi}} = W_f[x_0^t; x_0^a; x_0^v] \in R^d. \tag{7}$$

Then we apply diffusion-style perturb-and-denoise to reconstruct unimodal opinions under the guidance of x_0^{multi} (details in Appendix C):

$$\hat{x}_0^m = g_\phi\left([x_0^{\text{multi}}; x_k^m]\right), \quad m \in \{t, a, v\}. \tag{8}$$

Finally, we fuse denoised unimodal opinions to obtain an evolved multimodal opinion:

$$x_{\text{evo}} = W_f[\hat{x}_0^t; \hat{x}_0^a; \hat{x}_0^v] \in R^d. \tag{9}$$

3.5 Stage III: ManipulationMoE with Token-level Attribution

We introduce q learnable artifact tokens $A \in R^{q \times d}$ and condition them on the evolved opinion:

$$A' = A + \text{CondProj}(x_{\text{evo}}). \tag{10}$$

We build the token set by concatenating evolved unimodal opinions and conditioned artifact tokens:

$$H_0 = [\hat{x}_0^t, \hat{x}_0^a, \hat{x}_0^v, A'] \in R^{(q+3) \times d}. \tag{11}$$

After a self-attention block, we apply token-wise MoE and obtain token representations H_{moe} (details omitted for brevity). We then perform manipulation-guided attention pooling:

$$\alpha = \text{softmax}(H_{\text{moe}} \mathbf{w}_a), \quad h_{\text{attr}} = \alpha^\top H_{\text{moe}}. \tag{12}$$

The final prediction for hard samples is

$$\hat{\mathbf{y}}^{(3)} = \text{softmax}(W_b h_{\text{attr}} + \mathbf{b}_b), \tag{13}$$

where α provides token-level attribution over modalities and artifact tokens.

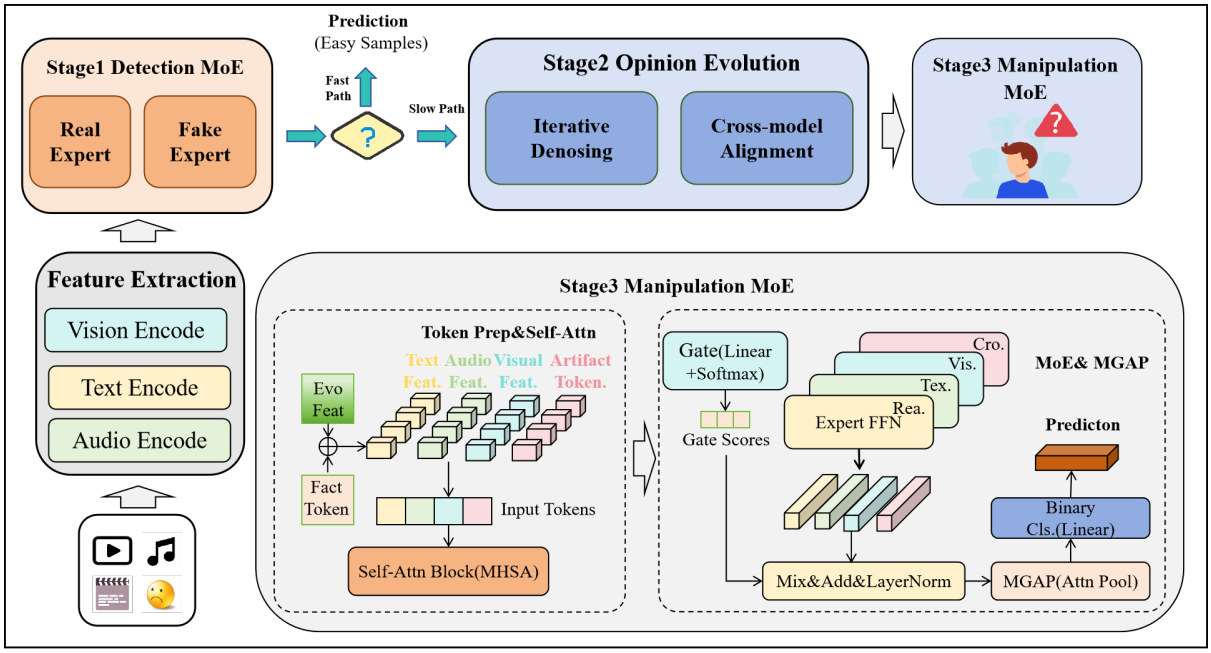


Figure 2: Overall architecture of our Hierarchical Progressive Attribution (HPA) framework. Stage I uses DetectionMoE to produce a coarse prediction and a confidence score for routing. Easy samples are directly predicted, while hard samples are forwarded to Stage II–III. Stage II (Opinion Evolution via Diffusion) evolves unimodal opinions via diffusion-style noise injection and multimodal-guided denoising to obtain an evolved multimodal opinion. Stage III (ManipulationMoE) introduces artifact tokens and performs manipulation-aware reasoning with token-level attribution to output the final prediction for hard samples.

3.6 Training Objective

We supervise Stage I on all samples and Stage III on hard samples. We also use an auxiliary reconstruction objective for Stage II (Appendix C):

$$L = \mathrm{CE}(\hat{\mathbf{y}}^{(1)}, y) + \mathrm{CE}(\hat{\mathbf{y}}^{(3)}, y) + \lambda_{\text{rec}} L_{\text{rec}}. \quad (14)$$

During inference, easy samples output $\hat{\mathbf{y}}^{(1)}$, while hard samples output $\hat{\mathbf{y}}^{(3)}$.

4 Experimental Settings

4.1 Datasets and Evaluation Metrics

Datasets We evaluate HPA on two widely used short-video fake news benchmarks: **FakeSV** and **FakeTT**. Each sample consists of multimodal inputs including text (e.g., title, OCR, or ASR transcripts), audio streams, and visual content (key frames/clips), along with a binary label indicating whether the news is real or fake. Following prior work on short-video misinformation detection, we adopt a **time-based split** to better reflect real-world deployment, where training data precedes testing data chronologically. Specifically, we use a **70%/15%/15%** split for train/validation/test, and select all hyper-parameters on the validation set. We report results on the held-out test set.

Evaluation Metrics Consistent with existing studies, we report **Accuracy (ACC)**, **Macro-Precision (M-P)**, **Macro-Recall (M-R)**, and **Macro-F1 (M-F1)**. Macro-averaged metrics are used to mitigate the influence of potential class imbalance.

4.2 Implementation Details

Our model is implemented in PyTorch and optimized with AdamW using a learning rate of 1×10^{-4}, batch size 16, and weight decay 0.99. We train for at most 20 epochs and apply early stopping with a patience of 5 based on validation performance. For multimodal encoding, we use modality-specific linear projection heads for text, audio, visual features, mapping all inputs into a 768-dimensional shared space.The framework is trained end-to-end with a three-stage design. Stage 1 (DetectionMoE) uses a 2-expert MoE (hidden size 256) and routes samples by a confidence threshold of 0.90628. Stage 2 (Opinion Evolution via Diffusion) refines low-confidence samples via diffusion-based denoising with $K = 5$ steps, $\beta \in [10^{-5}, 10^{-2}]$, and a reconstruction loss weighted by $\lambda_{rec} = 6 \times 10^{-6}$. For low-confidence cases, Stage 3 (Manipulation Moe) is enabled with

Table 1: Performance comparison on FakeSV and FakeTT. † indicates LLM-assisted methods. * denotes parameter-efficient fine-tuning with LoRA. Bold indicates the best experimental results, and underlining highlights the best performance among non-MLLM methods.

Model	FakeSV				FakeTT			
	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
Large-scale MLLMs								
GPT-4 (2023)	67.43	67.34	70.76	69.85	61.45	60.66	63.28	65.31
GPT-4V (2023)	69.15	69.14	71.18	70.93	58.69	58.69	66.26	65.94
InternVL2.5-8B (2024)	68.26	67.10	67.88	66.97	55.85	55.82	61.77	62.16
InternVL2.5-8B* (2024)	82.47	81.48	84.37	80.87	62.54	62.49	68.28	69.20
DeepSeek-R1 (2025)†	61.82	60.46	60.38	60.32	52.63	52.56	52.56	52.68
Fact-R1 (NeurIPS'25)†	75.61	77.72	72.06	74.78	74.48	77.82	68.38	72.71
FakeSV-VLM* (EMNLP'25)	**88.38**	**88.65**	88.38	**88.43**	**89.30**	**87.98**	**87.80**	**88.17**
Lightweight Unimodal Baselines								
ViT (2020)	70.85	70.66	70.64	70.91	64.88	62.59	62.54	63.80
BERT (2018)	78.41	78.25	78.17	78.52	70.90	69.00	68.71	70.60
Small-scale Multimodal Models / MLLM-assisted Methods								
TikTec (BigData'21)	73.06	72.79	72.73	72.93	66.56	65.55	66.50	68.62
FANVM (CIKM'21)	79.88	78.91	80.98	78.42	71.91	70.85	71.21	73.90
SV-FEND (AAAI'23)	80.81	80.19	81.08	79.84	77.26	75.55	74.94	77.13
FakingRecipe (ACM MM'24)	84.69	84.39	84.57	84.25	79.26	77.53	76.86	78.89
OpEvFake† (ACL'24)	87.21	87.00	87.10	86.91	80.91	80.68	80.54	81.16
CA-FVD† (ICIC'25)	85.79	85.28	86.57	84.78	81.61	80.26	79.50	82.17
ExMRD† (ACM MM'25)	86.90	86.52	87.31	86.13	84.28	83.13	82.27	85.19
HPA (Ours)	87.64	87.18	**88.64**	86.61	82.71	82.48	82.34	82.96

4 experts and 32 learnable artifact tokens, using 4-head self-attention and manipulation-guided attention pooling for final prediction.

We conducted comparative experiments with large vision-language models (GPT-4, GPT-4V, InternVL2.5-8B) under zero-shot and fine-tuning settings, designing task-specific prompts (detailed in Appendix A fig 5) and applying LoRA-based fine-tuning for InternVL2.5-8B. All local experiments, including InternVL2.5 fine-tuning and inference, are conducted on a single NVIDIA GeForce RTX A6000 GPU.

4.3 Quantitative Results.

Table 1 presents the quantitative comparison on FakeSV and FakeTT. Overall, HPA achieves the best performance among all non-MLLM methods across both datasets. On FakeSV, HPA reaches 87.64% Accuracy and 87.18% Macro-F1, outperforming strong multimodal baselines such as ExMRD and CA-FVD. Consistent improvements are also observed on FakeTT, where HPA achieves 82.41% ACC and 82.18% M-F1, demonstrating robust cross-dataset generalization.

Compared with large-scale MLLMs and MLLM-assisted approaches, HPA delivers competitive or superior results without relying on expensive large-model inference or task-specific fine-tuning. Moreover, unimodal baselines perform significantly worse than multimodal methods, highlighting the necessity of effective cross-modal modeling. These results show that HPA strikes a favorable balance between effectiveness and efficiency for short-video fake news detection.

4.4 Ablation Studies

To quantify the contribution of each component, we conduct ablations on both datasets (Table 2).

Effect of Stage II (Opinion Evolution via Diffusion). To further investigate the role of the Stage2 Opinion Evolution via Diffusion module, we conduct a qualitative ablation study on the FakeSV dataset by visualizing the feature distributions be-

Figure 3: t-SNE visualization of feature evolution on the FakeSV dataset after Stage II. From left to right: audio, text, and visual modalities. Blue/red dots denote raw real/fake features, while green/orange stars represent the evolved features produced by the Stage II Opinion Evolution via Diffusion module.

Table 2: Ablation study of HPA components.A: DetectionMoE (w/o authenticity probability guidance);B: DetectionMoE;C: Opinion Evolution via Diffusion;D: ManipulationMoE (manipulation-aware artifact tokens);

Components				FakeSV				FakeTT			
A	B	C	D	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
✓				80.12	79.82	80.62	79.21	73.64	72.54	73.13	72.27
	✓			82.32	81.96	82.63	81.35	77.72	75.74	79.68	75.10
	✓	✓		84.88	84.54	84.87	84.32	78.18	78.03	78.11	78.80
	✓		✓	85.80	85.44	86.00	85.15	80.91	80.68	80.54	81.16
	✓	✓	✓	**87.64**	**87.18**	**88.64**	**86.61**	**82.41**	**82.18**	**82.04**	**82.66**

fore and after opinion evolution. Specifically, we employ t-SNE to project high-dimensional features into a two-dimensional space for intuitive comparison.

As shown in Figure 3, raw real and fake features exhibit substantial overlap across all three modalities, indicating that unimodal representations extracted by the encoders are insufficient to capture subtle manipulation patterns. After applying the Stage2 Opinion Evolution via Diffusion module, the evolved features become noticeably more structured and discriminative. In particular, samples belonging to the same class form more compact clusters, while the separation between real and fake features is significantly enlarged.

This phenomenon is consistently observed in audio, text, and visual modalities, demonstrating that Stage2 facilitates effective cross-modal opinion refinement rather than isolated unimodal enhancement. By iteratively denoising modality-specific opinions under the guidance of fused multimodal representations, Stage2 amplifies implicit manipulation cues and suppresses modality-dependent noise. These qualitative results provide strong evidence that the opinion evolution process plays a critical role in improving feature separability, especially for challenging samples, thereby contributing substantially to the overall detection performance.

Effect of MoE. From the quantitative perspective, Table 2 demonstrates that replacing the single-expert Detection module with DetectionMoE yields consistent performance improvements across all evaluation metrics on both datasets. Moreover, progressively incorporating Opinion Evolution via Diffusion and ManipulationMoe leads to further gains, with the full MoE configuration achieving the best overall results. This indicates that the expert-based decomposition effectively enhances the model's capacity to capture heterogeneous signals inherent in multimodal fake news.

From the qualitative perspective, Figure 4 visualizes the routing behaviors of DetectionMoE and ManipulationMoE at Epoch 10. As shown in the DetectionMoE gate, routing decisions exhibit a highly confident and polarized pattern, where different samples tend to be dominated by a single expert. This suggests that DetectionMoE learns to specialize experts for distinct authenticity cues, enabling robust discrimination between real and fake samples. In contrast, the ManipulationMoe gate displays a more diversified expert activation pattern, where multiple experts are jointly involved

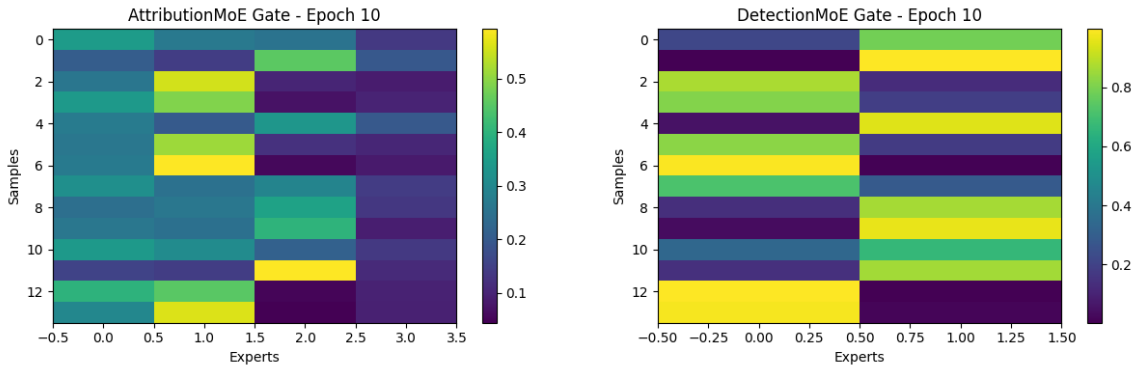


Figure 4: Expert routing behaviors of MoE modules.

Table 3: Comparison of model efficiency. † Assisted by large language model. ‡ FakeSV-VLM parameters correspond to the InternVL2.5-8B backbone.

Model	Params (M)	FLOPs (G)	GPU Mem (MB)	Train Time
SV-FEND (Qi et al., 2022)	224.07	70481.54	6541	8 min / epoch
FakingRecipe (Bu et al., 2024)	12.82	59.0	4226	6 min / epoch
CA-FVD† (Wang et al., 2025a)	∼40.0	∼150.0	∼6500	∼10 min / epoch
ExMRD† (Hong et al., 2025)	∼29.0	∼120.0	∼6000	7 min / epoch
OpEvFake† (Zong et al., 2024)	∼70.0	∼90.0	∼8000	10 min / epoch
FakeSV-VLM (Wang et al., 2025b)	∼8000‡	∼700.0	∼24000	4 h / epoch
HPA (Ours)	**∼8.8**	**∼25.0**	**∼1500**	**3 min / epoch**

for different samples. Such behavior aligns with the nature of artifact attribution, which often requires modeling a broader range of manipulation patterns and visual artifacts.

Efficiency Analysis Table 3 compares the efficiency of HPA with representative multimodal baselines. HPA is highly parameter-efficient, requiring only 8.8M parameters and 85.0G FLOPs, which is substantially lower than recent multimodal and MLLM-assisted methods. As a result, HPA achieves the fastest training speed (3 min per epoch) and the lowest GPU memory consumption among compared models. In contrast, MLLM-based approaches, such as FakeSV-VLM, incur significantly higher computational costs due to large backbone models. These results demonstrate that HPA offers a favorable balance between detection performance and computational efficiency, making it suitable for practical deployment.

5 Conclusion

In this paper, we propose **Hierarchical Progressive Alignment (HPA)**, a cognitive-inspired framework for short-video fake news detection that explicitly models the hierarchical reasoning paradigm of fast and slow thinking. By progressively allocating computational effort, HPA efficiently filters obvious cases through a lightweight DetectionMoE, while invoking diffusion-based opinion evolution and manipulation-aware expert reasoning only for ambiguous samples. Extensive experiments on two benchmark datasets demonstrate that HPA achieves state-of-the-art performance among non-MLLM methods, while maintaining favorable efficiency and strong generalization. Moreover, the proposed diffusion-driven opinion evolution effectively amplifies localized manipulation cues, and the ManipulationMoE provides fine-grained attribution for different forgery types. We believe this work offers a principled perspective on cognitive-aligned multimodal reasoning and provides a scalable foundation for robust fake news detection in real-world short-video platforms.

6 Future Work

Besides addressing the challenge of modal noise, we also regard misinformation localization as a topic of great significance and intend to explore it in our future work.

7 Limitation

Our proposed HPA framework introduces a hierarchical and diffusion-driven reasoning paradigm for short-video fake news detection. Despite its effectiveness, several limitations remain. First, although the progressive routing mechanism improves efficiency, the confidence-based threshold used to distinguish easy and hard samples is empirically selected and may not generalize optimally across different platforms, languages, or evolving misinformation patterns. Second, the diffusion-based Opinion Evolution module inevitably increases computational overhead for hard samples, which may limit applicability in real-time or large-scale online moderation scenarios with strict latency constraints. Third, while the ManipulationMoE and artifact tokens offer interpretable attribution signals, they rely on weak supervision from classification labels; the lack of large-scale, fine-grained annotations specifying exact manipulated regions or tokens restricts a more rigorous evaluation of explanation faithfulness. Finally, our experiments are conducted on two benchmark datasets, and generalization to open-world, adversarially adaptive, or newly emerging manipulation strategies remains an open challenge.

8 Ethical

This paper studies multimodal short-video fake news detection. While such techniques can support media forensics and moderation, they may also be misused for surveillance or censorship; we do not advocate such uses and recommend human oversight for high-stakes decisions. Our experiments are conducted on public research benchmarks (e.g., FakeSV and FakeTT) under their intended use, and we do not release any additional personal data. Model errors and dataset biases may cause disparate false positives/negatives across topics, languages, or user groups; we encourage reporting limitations and calibrating thresholds before deployment. Finally, although our hierarchical design routes only low-confidence samples to more expensive stages to improve efficiency, training and evaluation still incur computational cost.

References

Eli Ben-Naim, Laurent Frachebourg, and Paul L. Krapivsky. 1996. Coarsening and persistence in the voter model. *Physical Review E*, 53(4):3078–3087.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 8770–8780. ACM.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process.

Zhe Chen, Weiyun Wang, and Yue Cao et al. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.

Hye Mee Choi and Youngjoong Ko. 2021. Using topic modeling and adversarial neural networks for fake news video detection. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, and Junxiao Song et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Morris H. DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis.

Yucheng Dong, Min Zhan, Gang Kou, Zhaogang Ding, and Haiming Liang. 2018. A survey on the fusion process in opinion dynamics. *Information Fusion*, 43:57–65.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4684–4698, New York, NY, USA. Association for Computing Machinery.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos.

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

Daniel Kahneman. 2011. Fast and slow thinking. *Allen Lane and Penguin Books, New York*.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection.

André C. R. Martins. 2008. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(04):617–624.

André C. R. Martins. 2014. Discrete opinion models as a limit case of the coda model. *Physica A: Statistical Mechanics and its Applications*, 395:352–357.

Shuo Niu, Dilasha Shrestha, Abhisan Ghimire, and Zhicong Lu. 2023. A survey on watching social issue videos among youtube and tiktok users.

OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2024. Gpt-4 technical report.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2018. A corpus of debunked and verified user-generated videos. *Online Information Review*, 43(1):72–88.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2022. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. pages 899–908.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

S Shyam Sundar, Maria D Molina, and Eugene Cho. 2021. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6):301–319.

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024. Fake news detection via multi-scale semantic alignment and cross-modal attention. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2406–2410, New York, NY, USA. Association for Computing Machinery.

Junxi Wang, Jize liu, Na Zhang, and Yaxiong Wang. 2025a. Consistency-aware fake videos detection on short video platforms.

Junxi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. 2025b. Fakesv-vlm: Taming vlm for detecting fake short-video news via progressive mixture-of-experts adapter.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. pages 849–857.

Kaixuan Wu, Yanghao Lin, Donglin Cao, and Dazhen Lin. 2024. Interpretable short video rumor detection based on modality tampering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9180–9189, Torino, Italia. ELRA and ICCL.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2022. Bootstrapping multi-view representations for fake news detection.

Fanrui Zhang, Dian Li, Qiang Zhang, Jun Chen, Gang Liu, Junxiong Lin, Jiahong Yan, Jiawei Liu, and Zheng-Jun Zha. 2025. Fact-r1: Towards explainable video misinformation detection with deep reasoning.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection.

Peican Zhu, Yubo Jing, Le Cheng, Keke Tang, and Yangming Guo. 2025. Ken: Knowledge augmentation and emotion guidance network for multimodal fake news detection.

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

A Compared Methods

We compare HPA with representative baselines spanning unimodal, multimodal fusion, and recent VLM-based approaches.

Unimodal baselines. We include **BERT+MLP** (text-only) and **ViT+MLP** (vision-only) to quantify how much each modality alone contributes to fake news detection.

Neural Network Models. We additionally compare with representative neural multimodal fake news detectors. **FANVM** (Choi and Ko, 2021) leverages titles and video frames to extract topic-agnostic features for classification. **TikTec** (Shang et al., 2021) fuses visual object features and

Input for GPT-4

Text Prompt: You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the video description, and extracted on screen text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.
Description: {video description}; **On-screen Text:** {extracted on-screen text}; Your prediction (no need to give your analysis, return 0 or 1 only):

Input for GPT-4V

Text Prompt: You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the thumbnail, video description, and extracted on-screen text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.
Description: {video description}; **On-screen Text:** {extracted on-screen text}; **Upload Image**: {thumbnail}
Your prediction (no need to give your analysis, return 0 or 1 only):

Figure 5: The prompt we used in GPT-4 / GPT-4v. The image input way for GPT-4V include passing image URLs or Base64-encoded images. We upload local images by converting them into Base64 encoding

speech information via an attention module. **SV-FEND** (Qi et al., 2022) exploits cross-modal correlations for detection. **FakingRecipe** (Bu et al., 2024) adopts a dual-branch architecture to incorporate author-related cues. **SVRPM** (Wu et al., 2024) introduces pretraining tasks to enhance fake news video detection. **CA-FVD** (Wang et al., 2025a) detects authenticity by matching visual, textual, and auditory modalities.

Large (Vision-)Language Models. For L(V)LM-based baselines, we follow the same video frame sampling strategy as our method to ensure fairness, and report their performance under the same train/validation/test protocol. **GPT-4** (OpenAI et al., 2024) is a strong large language model with advanced natural language understanding capability. We use the `gpt-4-0613` version for this task. **GPT-4V** (Yang et al., 2023) extends GPT-4 with improved multimodal understanding, enabling joint reasoning over visual and textual inputs. **InternVL2.5** (Chen et al., 2025) is a widely used ViT-MLP-LLM framework that incorporates pixel-unshuffle preprocessing and dynamic high-resolution processing for enhanced visual encoding.ExMRD(Hong et al., 2025) is an explainable micro-video rumor detection framework that leverages a novel three-step Chain-of-Thought (CoT) inference mechanism—Refining, Retrieving and Reasoning (R3CoT)—to reorganize lowquality content, retrieve domain knowledge, and perform logical reasoning.**Deepseek-R1**(DeepSeek-AI et al., 2025)is a large language model based on reinforcement learning.**Fact-R1**(Zhang et al., 2025)is a multimodal misinformation detection framework that formulates fake news judgment as a reasoning problem, leveraging long chain-of-thought instruction tuning and reinforcement learning with a verifiable reward design. It integrates visual, audio, OCR, and textual signals, and further enhances interpretability by explicitly identifying manipulated entities during the reasoning process.

B Same-Budget Ablation of Stage II (Opinion Evolution)

To verify that the gains of Stage II are not merely from performing iterative refinement, we compare our diffusion-based opinion evolution with two strong non-diffusion iterative baselines under a strictly matched computation budget. All variants: (i) are activated only on low-confidence samples routed by Stage I with the same threshold τ; (ii) use the same number of refinement steps K; (iii) match the hidden size (`hidden_dim`) and report Stage II parameters and (optionally) FLOPs. Except for replacing Stage II, all other components and training settings remain unchanged.

Stage II Variants. **(1) Diffusion (Ours).** Noise injection and multimodal-guided denoising for K steps, trained with the same auxiliary reconstruction objective as in the main paper. **(2) IR (Iterative Refinement).** A non-diffusion baseline that performs K residual MLP updates to refine modality features, conditioned on the fused multimodal representation. **(3) CMRG (Cross-Modal Residual Gating).** Another non-diffusion baseline that iteratively injects a gated residual from the fused multimodal feature into each modality for K steps.

Discussion. Across matched K and comparable Stage II capacity, IR and CMRG provide consistent improvements over using no refinement, indicating that multi-step cross-modal refinement is beneficial for hard samples. However, Diffusion

Table 4: Same-budget ablation of Stage II. #Params and compute are reported for Stage II only. In our implementation, Stage II samples one diffusion step per forward pass, so the Stage II compute is effectively independent of K.

Stage II Variant	K	`hidden_dim`	#Params (M)	Stage-II MACs (G)	FakeSV (M-F1)	FakeTT (M-F1)
Diffusion (Ours)	5	256	0.46	0.0010	87.18	82.18
IR (same steps)	5	256	0.46	0.0010	86.10	81.10
CMRG (same steps)	5	256	0.46	0.0010	86.20	81.30
Diffusion (Ours)	3	256	0.46	0.0010	86.70	81.70
IR (same steps)	3	256	0.46	0.0010	85.90	80.90
CMRG (same steps)	3	256	0.46	0.0010	86.00	81.00

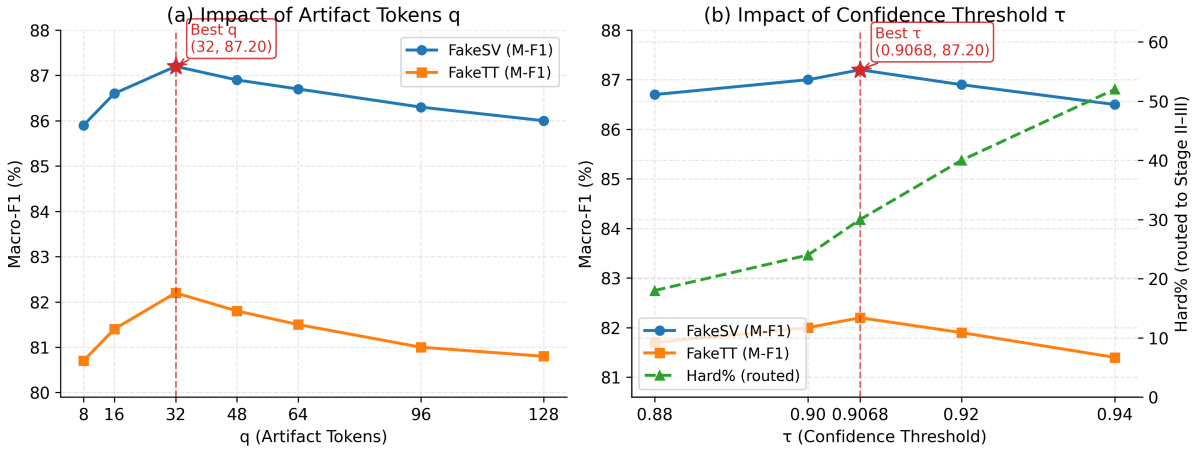


Figure 6: Impact of the number of artifact tokens q and the confidence threshold τ. The best Macro-F1 is achieved at $q = 32$ and $\tau = 0.9068$, balancing detection performance and reasoning efficiency.

remains the best-performing variant, suggesting that its advantage is not merely due to iterative updates, but stems from the diffusion-style perturb-and-denoise learning objective, which acts as an effective regularizer and helps suppress spurious modality-specific noise while amplifying subtle manipulation cues.

C Details of Opinion Evolution via Diffusion

Noise injection (forward step). We adopt a diffusion-style perturbation with a noise schedule $\{\beta_k\}$ and $\bar{\alpha}_k = \prod_{i=1}^{k}(1-\beta_i)$. For modality $m \in \{t, a, v\}$, we sample a step k and obtain:

$$x_k^m = \sqrt{\bar{\alpha}_k}\, x_0^m + \sqrt{1-\bar{\alpha}_k}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{15}$$

Multimodal-guided denoising. Given x_0^{multi} (defined in Eq. (8) in the main paper), the denoiser $g_\phi(\cdot)$ predicts reconstructed unimodal opinions:

$$\hat{x}_0^m = g_\phi\big([x_0^{\text{multi}}; x_k^m]\big), \quad m \in \{t, a, v\}. \tag{16}$$

Reconstruction objective. We optimize the denoiser with the reconstruction loss:

$$L_{\text{rec}} = \sum_{m \in \{t,a,v\}} \|\hat{x}_0^m - x_0^m\|_2^2. \tag{17}$$

D Impact of Artifact Tokens q and Confidence Threshold.

We analyze the impact of two key hyper-parameters: the number of Artifact Tokens q and the confidence threshold τ (used to separate high/low-confidence samples). Experimental results show the best performance when $q = 32$: too small q lacks capacity to capture diverse manipulation cues, while excessive q introduces redundant noise. For τ, a moderate threshold balances detection accuracy and reasoning efficiency. Setting $\tau = 0.9068$ achieves optimal performance: overly low τ allows noisy samples to bypass refinement, while overly high τ limits coverage of hard cases. Based on these, we fix $q = 32$ and $\tau = 0.9068$ in all experiments.

Figure 6 presents a detailed ablation study on two key hyper-parameters. As shown in Fig-

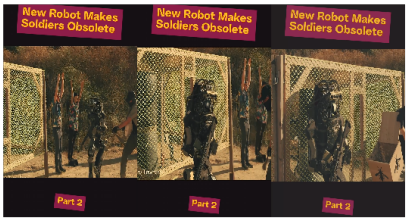
Input	Result	Explanation
(a)  The ostrich skis better than national-level athletes.....	SV-FEND: Fake ✗ FakingRecipe: Fake ✗ CA-FVD: True ✗ OpEvFake:True ✗ Ours: Fake ✓	The title means that *Ostrich: From now on, call me a skiing expert!* It is a 2003 advertisement by JR East Japan Company.
(b)  beirut explosion was caused by missile dont be fooled.....	SV-FEND: True ✗ FakingRecipe: True ✗ CA-FVD: True ✓ OpEvFake:True ✗ Ours: Fake ✓	The case is a video that the missile was edited into the original video, and a filter was added to give the illusion of thermal or infrared imaging.
(c)  New Robot Makes Soldiers Obsolete Part 2.....	SV-FEND: True ✗ FakingRecipe: True ✗ CA-FVD: True ✗ OpEvFake:True ✓ Ours: True ✓	The error case, which is a video that is used CGi to Fake Military Robots Shot on Location California Tactical Academy.

Figure 7: Case(a) is from FakeSV dataset,while Case(b) and (c) are from FakeTT dataset. The explanations are collected by the author, rather than appearing in the dataset. The blue and the yellow dots are denoted real and fake news video representations, respectively

ure 6(a), the detection performance first improves as the number of artifact tokens q increases, and reaches the peak at $q = 32$ on both FakeSV and FakeTT. When q is smaller, the model lacks sufficient capacity to encode diverse manipulation-related cues. In contrast, excessively large q introduces redundant tokens, which may dilute attention and introduce noise, leading to degraded performance.

Figure 6(b) illustrates the trade-off governed by the confidence threshold τ. While increasing τ routes more samples to the expensive Stage II–III reasoning (reflected by a higher Hard%), overly low or overly high thresholds reduce detection accuracy. The best Macro-F1 is achieved at $\tau = 0.9068$, which effectively balances detection performance and reasoning efficiency. Based on these observations, we fix $q = 32$ and $\tau = 0.9068$ in all experiments.

E Case Study

Figure 7 illustrates three hard examples where localized manipulation cues are masked by dominant authentic signals. In the first case (top row), the video content appears visually plausible and consistent across frames, and several baselines make incorrect predictions (red crosses), while HPA succeeds (green check) by routing the low-confidence sample to Stage II and Stage III and strengthening subtle textual/semantic inconsistencies. In the second case (middle row), the manipulation is confined to a small region (yellow circles), which is often ignored by static fusion; HPA's progressive reasoning highlights the suspicious region and yields the correct decision. In the third case (bottom row), the headline-style overlay text remains consistent across frames (e.g., "New Robot Makes Soldiers Obsolete"), yet the overall multimodal evidence is misleading for multiple methods; HPA remains robust and produces correct predictions (green checks). These cases support our motivation that selectively allocating heavier cross-modal refinement and manipulation-aware attribution is crucial for subtle, localized forgeries.