Exposing Cross-Modal Consistency for Fake News Detection in Short-Form Videos

Anonymous ACL submission

Abstract

Short-form video platforms are major channels for news but also fertile ground for multimodal misinformation where each modality appears plausible alone yet cross-modal relationships are subtly inconsistent, like mismatched visuals and captions. On two benchmark datasets, FakeSV (Chinese) and FakeTT (English), we observe a clear asymmetry: real videos exhibit high text–visual but moderate text–audio consistency, while fake videos show the opposite pattern. Moreover, a single global consistency score forms an interpretable axis along which fake probability and prediction errors vary smoothly. Motivated by these observations, we present **MAGIC**^{3} (Modal-Adversarial Gated Interaction and Consistency-Centric Classifier), a detector that explicitly models and exposes cross-tri-modal consistency signals at multiple granularities. MAGIC^3 combines explicit pairwise and global consistency modeling with token- and frame-level consistency signals derived from cross-modal attention, incorporates multi-style LLM rewrites to obtain style-robust text representations, and employs an uncertainty-aware classifier for selective VLM routing. Using pre-extracted features, MAGIC^3 consistently outperforms the strongest non-VLM baselines on FakeSV and FakeTT. While matching VLM-level accuracy, the two-stage system achieves 18–27× higher throughput and 93% VRAM savings, offering a *strong cost–performance tradeoff*.

1 Introduction

Short-form video platforms such as TikTok, Douyin, YouTube Shorts, and Kuaishou have reshaped how people consume news and public information (Bu et al., 2023). Given the sheer volume of uploads, detection systems must be not only accurate but also computationally efficient. Rich multimodal signals and highly optimized recommendation systems allow fake news *short videos* to attract massive attention before professional fact-checking becomes available, amplifying societal risks (Vosoughi et al., 2018; Zhou and Zafarani, 2020). Compared to text-only fake news (Devlin et al., 2019; Shu et al., 2019; Tian et al., 2025), short news videos are harder to analyze: manipulations may occur in any modality and at any granularity, and successful fakes rarely contain glaring artifacts in a single stream. Instead, they often exploit *cross-modal inconsistencies*, where each modality looks plausible alone but the joint message is misleading—as illustrated in Figure 1.

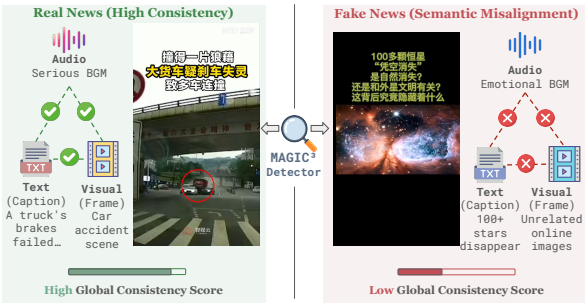

Figure 1: **Illustration of cross-modal consistency patterns.** In real news short videos, text, visuals, and audio are contextually aligned (Consistent). In fake news, a "semantic gap" often exists between the sensational claims (text/audio) and the actual visual content. MAGIC^3 acts as a consistency lens to quantify these multimodal relationships for fake news detection.

A consistency-centric view. We adopt a task-oriented taxonomy of cross-modal inconsistency in short-form news videos, covering entity/context, event semantics, affect, temporal audio–visual misalignment, and style/pragmatics, with detailed definitions and examples provided in Appendix A. Throughout this paper, we use "consistency" to denote *learned, task-oriented compatibility signals in feature space*, rather than factual correctness with respect to external world knowledge. Instead of designing yet another heavy backbone, we build a

lightweight, interpretable *consistency lens* that sits on top of any multimodal feature extractor and is tailored to fake news short videos.

Key findings. Grounded in this consistency-centric view, our analysis reveals four key insights: (i) *Asymmetric consistency*: real videos show high text–visual but moderate text–audio consistency, while fake videos flip this pattern; (ii) *Interpretable axis*: global consistency strongly correlates with prediction difficulty, with errors clustering at intermediate values; (iii) *Efficient routing*: combining consistency with uncertainty enables routing only 25% of samples to heavyweight VLMs while *surpassing* VLM-only accuracy; (iv) *Style robustness*: multi-style LLM rewrites improve robustness, with fake videos showing higher consistency variance.

Our method. We introduce **MAGIC**^{3}, a feature-level detector designed around explicit consistency modeling. Given frozen text, visual, and audio features (and a few LLM-rewritten texts), MAGIC^3 (i) computes pairwise and global consistency scores through the Cross-Modal Consistency Gate (CMCG), (ii) derives token- and frame-level *consistency fields* from cross-modal attention, (iii) builds style-robust text representations in an Adversarial-Aware Rewrite Fusion (AARF) module, (iv) performs lightweight hierarchical multimodal fusion with a temporal audio–visual inconsistency score (TCMI), and (v) exposes calibrated uncertainty estimates that drive a two-stage pipeline where expensive VLM-based detectors are invoked only when needed.

Contributions. We make three contributions: (i) To our knowledge, this is the first work on fake news short videos that makes tri-modal (text–visual–audio) consistency an *explicit*, multi-granularity output: pairwise and global scalar scores, plus token- and frame-level consistency fields that reveal text–visual vs. text–audio asymmetries and an interpretable consistency–difficulty axis. (ii) MAGIC^3, a detector that builds on these explicit consistency signals together with calibrated uncertainty estimates, providing an interpretable "consistency lens" for short-form news videos. (iii) A consistency- and uncertainty-driven two-stage VLM routing strategy that reaches state-of-the-art accuracy on FakeSV and FakeTT while offering 18–27× higher throughput, offering substantially better cost–performance tradeoff.

2 Related Work

Multimodal fake news on image–text pairs. Prior work extends fake news detection to image–text posts by jointly modeling captions and images (Zhou et al., 2020; Müller-Budack et al., 2020; Li et al., 2025). Event-adversarial methods such as EANN (Wang et al., 2018) learn event-invariant multimodal features to improve generalization, while similarity-aware approaches like SAFE (Zhou et al., 2020) encode image–text agreement or mismatch to flag manipulations. These methods motivate our focus on explicit cross-modal consistency, but they mainly target article- or post-level content rather than short videos with audio.

Short news video detection. Existing fake news detectors on short videos fall into three families. Feature-based multimodal models use strong single-modality encoders with late fusion like FakeSV Qi et al., 2023, but weakly exploit cross-modal interactions. Deep fusion architectures such as TikTec, FANVM, SVFEND, FakingRecipe, and KDSGAT-FNVD (Shang et al., 2021; Choi and Ko, 2021; Qi et al., 2023; Bu et al., 2024; Huang et al., 2025) learn powerful joint representations and *implicitly* reason about consistency, but do not expose where inconsistencies arise. More recent systems, like CA-FVD (Wang et al., 2025b), GLFE-SVFD (Wang et al., 2025a), and SCMG-FND (Yang et al., 2025), leverage consistency-aware training, global–local feature enhancement, or LLM-based semantic credibility scores to better capture multimodal relationships, yet their consistency cues remain coarse video-level signals or internal features. In contrast, we focus on short news videos with three modalities and turn multi-granularity consistency—pairwise/global scores, token-/frame-level consistency fields, and temporal audio–visual inconsistency—into explicit outputs that also drive an uncertainty-aware VLM routing.

3 Methodology

3.1 Problem Formulation

In this work, we focus on short-form news videos, as they constitute a dominant and practically important modality for online news consumption and multimodal misinformation. Both benchmarks we study, FakeSV and FakeTT, are constructed specifically from short videos on real-world platforms and reflect how users actually encounter news in today's media ecosystem. Our modeling choices,

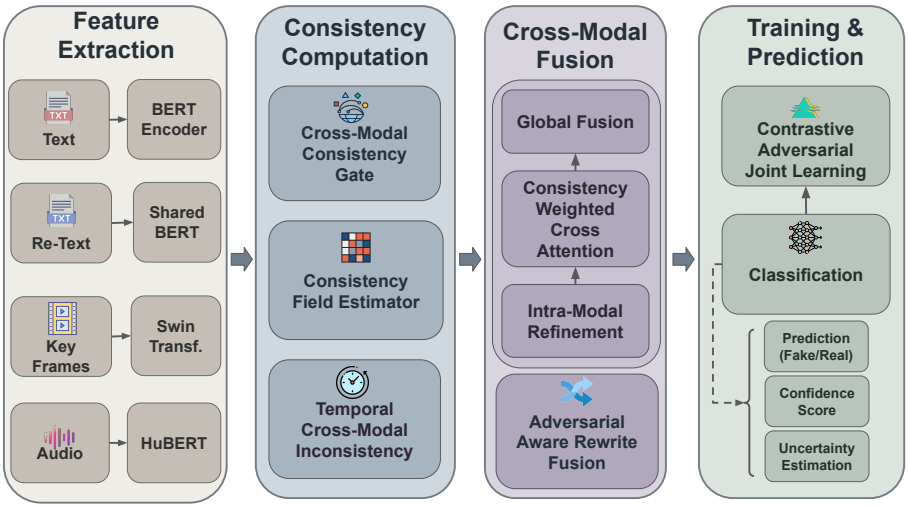

Figure 2: MAGIC^3 Overview. Frozen encoders provide text, visual, audio, and rewrite features. The Cross-Modal Consistency Gate outputs pairwise and global consistency scores; Consistency Field Estimator converts cross-modal attention into token- and frame-level consistency fields; Temporal Cross-Modal Inconsistency computes a temporal inconsistency score. Adversarial Aware Rewrite Fusion fuses original text with LLM rewrites into a style-robust representation; Hierarchical Multimodal Transformer performs hierarchical multimodal fusion; and the classifier outputs fake probability and uncertainty. The framework is trained via Contrastive Adversarial Joint Learning.

findings, and efficiency claims should therefore be interpreted within this practically important but intentionally scoped domain. Given a short news video $\mathcal{V}$ with associated text $\mathcal{T}$ (titles, subtitles, and descriptions), the task is to predict whether it is fake ($y = 1$) or real ($y = 0$).

3.2 Feature Extraction

Following prior work (Qi et al., 2023; Huang et al., 2025), we employ fixed pretrained encoders to extract feature-level representations. Specifically, we use BERT (Devlin et al., 2019) for encoding text titles and descriptions, Swin Transformer (Liu et al., 2021) for encoding visual keyframes, and ExHuBERT (Amiriparian et al., 2024) for audio emotion embeddings. This decoupled design lets MAGIC^3 leverage state-of-the-art representations without the burden of end-to-end fine-tuning. To enhance robustness, we also generate three LLM rewrites of each text (neutral, formal, sensational) and encode them with the same text encoder.

3.3 Consistency Computation

This module computes multi-granularity consistency signals to expose cross-modal discrepancies at different abstraction levels. We hypothesize that fake news videos exhibit inconsistencies in diverse ways: some are globally incoherent (e.g., unrelated subtitles and visuals), while others contain subtle, localized manipulations (e.g., a specific phrase contradicting a visual frame) or temporal artifacts. To capture this spectrum, we design three complementary mechanisms: (1) a global gate for high-level filtering, (2) fine-grained consistency fields for localized grounding, and (3) temporal inconsistency scores to detect synchronization artifacts. This multi-view approach ensures that MAGIC^3 remains sensitive to both cheapfakes and sophisticated manipulations.

Cross-Modal Consistency Gate (CMCG). As shown in Figure 2, this module pools modality features into vectors $\bar{\mathbf{h}}_{\text{text}}$, $\bar{\mathbf{h}}_{\text{vis}}$, and $\bar{\mathbf{h}}_{\text{aud}}$. It then computes scalar consistency scores for each pair (m_i, m_j) via a gated mechanism:

$$c_{ij} = \sigma(\mathbf{w}_{ij}^{\top}[\bar{\mathbf{h}}_{m_i}; \bar{\mathbf{h}}_{m_j}] + b_{ij}). \quad (1)$$

These pairwise scores form a vector $\mathbf{c} = [c_{\text{tv}}, c_{\text{ta}}, c_{\text{va}}]$, which provides a compact summary of inter-modal agreement. Finally, they are compressed into a single global consistency scalar:

$$c_{\text{global}} = \sigma(\text{MLP}_c(\mathbf{c})). \quad (2)$$

This global score c_{global} provides a high-level measure of the video's overall coherence and modulates cross-modal attention. Throughout, we use "consistency" to denote learned, task-oriented compatibility proxies computed in frozen encoder feature spaces, keeping the consistency lens lightweight and decoupled from heavy backbones.

Consistency Field Estimator (CFE). It converts cross-modal attention matrices into per-token and per-frame "consistency fields", localizing *where* a video is inconsistent. Given a cross-attention matrix $\mathbf{A}_{m_i \leftarrow m_j} \in \mathbb{R}^{L_{m_i} \times L_{m_j}}$ from the cross-modal transformer, we define:

$$F_{ij}^{(m_i)}(t) = \max_k \mathbf{A}_{m_i \leftarrow m_j}(t,k), \quad t = 1, \ldots, L_{m_i}. \tag{3}$$

Intuitively, $F_{ij}^{(m_i)}(t) \in [0,1]$ reflects how strongly token t in modality m_i finds supporting evidence in modality m_j. Tokens that attend strongly and with high focus receive high scores; tokens that attend diffusely or weakly receive low scores. These fields support heatmap visualizations and localized analysis, and visualizations of some examples of FakeSV and FakeTT are shown in Figure 6 in Appendix B.

Temporal Cross-Modal Inconsistency (TCMI). In parallel, we align audio and visual features along time to capture coarse mismatches. We resample both modalities to length T' and compute per-timestep distances:

$$d_t = \left\| \tilde{\mathbf{h}}_{\text{vis},t} - \text{Proj}_{\text{av}}(\tilde{\mathbf{h}}_{\text{aud},t}) \right\|_2, \tag{4}$$

which are summarized into $\mathbf{s}_{\text{temp}} = [\text{mean}(\mathbf{d}); \text{var}(\mathbf{d}); \max(\mathbf{d})]$ and mapped to:

$$c_{\text{temp}} = 1 - \sigma\big(\text{MLP}_{\text{temp}}(\mathbf{s}_{\text{temp}})\big). \tag{5}$$

This captures coarse audio–visual mismatches inspired by deepfake detectors (Gao et al., 2024).

3.4 Cross-Modal Fusion

Adversarial-Aware Rewrite Fusion (AARF). The Multi-view Adversarial-Aware Rewrite Fusion module treats LLM rewrites as semantic-preserving but style-altering views (like in neutral, formal, and sensational styles). Let $\mathbf{h}_{\text{orig}} = \text{Pool}(\mathbf{H}_{\text{text}})$ be the original text and $\mathbf{h}_{\text{rew}}^{(v)}$ the v-th rewrite. We compute gated weights:

$$q^{(v)} = \sigma\big(\text{MLP}_q[\mathbf{h}_{\text{orig}}; \mathbf{h}_{\text{rew}}^{(v)}]\big), \tag{6}$$

where $q^{(v)}$ estimates rewrite quality. A gating MLP produces fusion logits from all views, and we enforce a minimum weight $\alpha_{\min}$ on the original text:

$$\mathbf{h}_{\text{text}}^{\text{fuse}} = \alpha_0 \, \mathbf{h}_{\text{orig}} + \sum_{v=1}^{V} \alpha_v \, \text{Proj}\big(\mathbf{h}_{\text{rew}}^{(v)}\big), \tag{7}$$

where $\alpha_0 \geq \alpha_{\min}$ and $\sum_v \alpha_v = 1$, and $\text{Proj}(\cdot)$ is a learned linear projection that aligns the rewrite embeddings with the original text embedding dimension. Contrastive losses pull together representations of rewrites from the same video and push apart those from different videos, encouraging style-invariant embeddings.

Hierarchical Multimodal Transformer (HMT). This component operates in three stages. (i) *Intra-modal refinement*: for each modality $m \in \{\text{text}, \text{visual}, \text{audio}\}$, we apply a modality-specific Transformer encoder with self-attention, $\mathbf{H}_m^{(A)} = \text{TransEnc}_m(\mathbf{H}_m)$, to refine token or frame sequences within that modality. (ii) *Consistency-weighted cross-attention*: we then add cross-attention blocks where queries from one modality attend to keys/values from another modality, and aggregate the resulting messages using weights derived from the pairwise consistency vector $\mathbf{c}$. The cross-attention matrices in this stage are also used by CFE to construct token- and frame-level consistency fields. (iii) *Global aggregation*: finally, we pool the refined representations from all modalities (and the fused text) and feed them into a small global fusion layer to obtain a single video-level representation $\mathbf{h}_{\text{global}}$ that is passed to the classifier.

3.5 Training and Prediction

We train MAGIC^3 using a Contrastive–Adversarial Joint Learning (CAJL) objective to jointly optimize classification accuracy and the reliability of consistency and uncertainty signals for prediction and routing. The primary supervised loss is standard binary cross-entropy:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N} \sum_i \Big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big). \tag{8}$$

To improve robustness and alignment, we employ InfoNCE-style contrastive losses for intra- and cross-modal alignment ($\mathcal{L}_{\text{intra}}, \mathcal{L}_{\text{cross}}$). For a batch of N pairs $(\mathbf{u}_i, \mathbf{v}_i)$, the loss is:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N} \sum_i \log \frac{\exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{u}_i, \mathbf{v}_j)/\tau)}. \tag{9}$$

To improve prediction robustness, we also introduce random noise perturbations to the hidden representations during training and minimize the KL-divergence between predictions on clean and perturbed features, formulated as an adversarial consistency regularizer $\mathcal{L}_{\text{adv}}$. The final composite objective combines these terms with a consistency

regularizer $\mathcal{L}_{\text{reg}}$ that aligns global and local scores:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \quad (10)$$

where λ are trade-off hyperparameters. More details on all terms are provided in Appendix C.

Classifier. The final classifier produces three distinct signals: (1) Prediction: The binary classification probability $\hat{y}_{fake}$ indicating the likelihood of the video being fake. (2) Confidence Score: Derived from the softmax output as conf $= \max(\hat{y}, 1-\hat{y})$, representing the model's confidence in its chosen class. (3) Uncertainty Estimation: A composite metric combining predictive entropy $u_{\text{ent}} = -\sum p \log p$ and the learned auxiliary scalar $\hat{u}$ (trained to regress the error margin). These signals provide a robust basis for the subsequent routing decision, allowing the system to distinguish between confident errors and genuine ambiguity.

Two-stage Routing. During inference, we employ a routing function $\mathcal{R}(x)$ to decide whether to trust MAGIC^3's prediction or escalate to a heavyweight VLM. Unlike simple confidence thresholding, we leverage our multi-dimensional outputs to define effective routing strategies (detailed in Appendix D.4): (1) *Uncertainty-based*: Routes samples with high predictive entropy ($u_{\text{ent}} > \tau$), prioritizing cases where the model is information-theoretically uncertain. (2) *Difficulty-based*: Uses a composite difficulty score $d = \text{Norm}(u_{\text{ent}}) + (1 - c_{\text{global}}) + (1 - \text{conf})$ to route samples that are simultaneously uncertain, inconsistent, and low-confidence. By filtering out the ∼75% of "easy" cases that MAGIC^3 handles correctly, we achieve VLM-level accuracy at a fraction of the cost.

4 Experiments

4.1 Experimental Setup

We evaluate MAGIC^3 on the FakeSV and FakeTT benchmarks for short video fake news detection (Qi et al., 2023; Huang et al., 2025). We follow the chronological split protocol of Huang et al. (2025), using 70%/15%/15% of videos for training, validation, and testing. Dataset statistics are summarized in Table 10 and detailed further in Appendix E.

Text, visual, and audio features are extracted as described in Section 3.2 using fixed pretrained encoders. All trainable parameters reside in the MAGIC^3 detector head (CMCG, CFE, AARF, HMT+TCMI, and the classifier). We use AdamW with learning rate 5×10^{-5}, batch size 32, cosine decay with 10% warm-up, hidden size 256, and 8 attention heads. LLM rewrites are generated offline using DeepSeek-V3.2 with three style prompts (neutral, formal, sensational); detailed prompts are listed in Appendix F. Loss weights and hyperparameters are tuned on the validation set (Table 11 in Appendix G). Once video/audio/text features are pre-extracted, the detector head reaches roughly ∼680 feature-level samples per second with batch size 128 on a single RTX 5090 GPU (about 1.5 ms/sample and ∼1.67 GB peak GPU memory on cached features); end-to-end latency is dominated by decoding and feature extraction, which can be amortized via caching and asynchronous preprocessing. We have more efficiency and deployment discussion in Appendix H as well.

4.2 Baselines

We compare MAGIC^3 against three categories of baselines. (i) *Zero-shot VLMs*: GPT-4o-mini, GPT-4.1-mini, Qwen2.5-VL (Bai et al., 2025), InternVL2.5 (Chen et al., 2024), and InternVL2.5-MPO (Wang et al., 2024b), evaluated by directly prompting them on the detection task. (ii) *Supervised multimodal methods*: HCFC-Hou (Bu et al., 2024), HCFC-Medina (Serrano et al., 2020) (hand-crafted feature classifiers), TikTec (Shang et al., 2021), FANVM (Choi and Ko, 2021), SVFEND (Qi et al., 2023), FakingRecipe (Bu et al., 2024), and KDSGAT-FNVD (Huang et al., 2025). (iii) *VLM-based detector*: FakeSV-VLM (Wang et al., 2025c), a fine-tuned VLM specifically designed for fake news video detection.

4.3 Main Results

Table 1 reports results on FakeTT and FakeSV. We compare MAGIC^3 with the baselines detailed in Section 4.2. For each dataset, we report accuracy, macro-F1, and per-class precision/recall. On FakeSV, MAGIC^3 reaches 86.71% accuracy and 86.84 macro-F1, outperforming KDSGAT-FNVD by 1.98 and 2.16 points, respectively, with balanced precision/recall across real and fake classes. On FakeTT, MAGIC^3 achieves 84.95% accuracy and 82.28 macro-F1, improving over KDSGAT-FNVD by 3.01 and 3.34 points. Zero-shot VLMs lag far behind: on FakeSV the best VLM trails MAGIC^3 by about 16 accuracy points, and on FakeTT the gap exceeds 20 points, underscoring the value of supervision and explicit consistency modeling. These gains also hold across macro-F1 and per-class precision/recall, particularly on the real class of FakeTT

Method	FakeSV						FakeTT					
	Acc	macF1	Real		Fake		Acc	macF1	Real		Fake	
			Pre	Rec	Pre	Rec			Pre	Rec	Pre	Rec
Zero-shot VLMs												
GPT-4o-mini	68.08	68.05	65.35	73.97	74.41	64.93	61.54	61.20	81.62	52.00	47.28	79.80
GPT-4.1-mini	70.30	70.25	68.12	73.53	73.10	68.21	49.16	48.54	77.42	48.00	37.50	71.40
Qwen2.5-VL	64.21	60.79	62.50	60.54	66.60	62.50	45.82	45.31	72.10	45.00	41.28	65.84
InternVL2.5	64.39	57.89	63.80	57.50	73.24	63.50	46.82	45.29	77.60	49.00	52.24	69.46
InternVL2.5-MPO	65.13	61.07	63.75	59.16	69.17	65.08	43.14	40.84	74.00	47.00	49.80	65.46
Supervised baselines												
HCFC-Hou	74.91	73.61	77.72	60.08	73.46	86.51	73.24	72.00	87.04	70.50	56.93	78.79
HCFC-Medina	76.38	75.83	74.77	69.75	77.50	81.58	62.54	62.23	84.92	53.50	46.24	80.81
TikTec	73.43	73.26	68.08	74.37	78.37	72.70	66.22	65.08	82.35	63.00	49.32	72.73
FANVM	79.52	78.81	80.98	69.75	78.64	87.17	71.57	70.21	85.28	69.50	55.15	75.76
SVFEND	80.88	80.17	74.53	83.61	85.82	77.63	77.14	75.63	87.91	76.33	62.33	78.79
FakingRecipe	82.87	83.52	**91.35**	71.00	80.67	**94.74**	79.26	77.87	89.66	78.00	64.80	81.82
KDSGAT-FNVD	84.73	84.68	82.23	83.61	87.00	85.86	81.94	78.94	92.68	76.00	64.44	87.88
Ours												
FakeSV-VLM	89.60	89.40	89.20	90.10	90.50	88.00	88.90	87.60	88.20	86.50	86.60	89.10
MAGIC^3 (ours)	86.71	86.84	86.52	83.61	87.50	89.80	84.95	82.28	**94.12**	80.00	68.99	89.90
MAGIC^3+VLM (ours)	**90.93**	**90.41**	90.85	**90.95**	**91.10**	88.90	**89.52**	**88.14**	89.70	**88.50**	**87.20**	**90.10**
End-to-End Efficiency												
FakeSV-VLM	18.67 samp/s @24.68 GB						30.25 samp/s @22.76 GB					
MAGIC^3 (ours)	**680.43 samp/s @1.67 GB**						**693.76 samp/s @1.67 GB**					
MAGIC^3+VLM (ours)	509.71 samp/s @24.68 GB						530.73 samp/s @22.76 GB					

Table 1: Main results on FakeSV and FakeTT. Best values in **bold**, second best underlined. MAGIC^3+VLM routes 25% of samples to VLM (Figure 5) while achieving comparable or better accuracy, offering the **best cost–performance tradeoff**: VLM-level or higher accuracy at 18–27× higher throughput than VLM-only.

where MAGIC^3 attains the highest precision.

When combined with a heavyweight VLM (FakeSV-VLM) in a two-stage pipeline, MAGIC^3+VLM makes further improvement and even surpasses the specialized VLM-only detector FakeSV-VLM on both FakeSV (90.93% vs. 89.60%) and FakeTT (89.52% vs. 88.90%), while still clearly outperforming all non-VLM baselines (Table 1). A detailed analysis of the routing policy and its superior cost–performance benefits is provided in Section 5.3.

4.4 Ablation Study

To assess the contribution of each component, we conduct ablations on both datasets (Table 2). Removing core consistency modules (CMCG, CFE, TCMI) consistently degrades performance, confirming that explicit consistency modeling is effective. AARF style-robust fusion impacts performance particularly on FakeTT, supporting the need for handling style shifts. CFE and TCMI also contribute to interpretability, driving the field visualizations and temporal inconsistency detection.

5 Analysis and Findings

Beyond the main results, we investigate the consistency signals to uncover deeper insights about how real and fake videos differ.

5.1 Asymmetric Cross-Modal Consistency

We first examine cross-modal consistency in aggregate for real vs. fake videos. Using CMCG, we compute four scalar scores per video: text–visual, text–audio, visual–audio, and global consistency. Figure 3 shows their distributions for FakeSV (left) and FakeTT (right). In both datasets, real videos exhibit high text–visual and moderate text–audio consistency, whereas fake videos show the reverse: text–audio consistency is very high, but text–visual consistency collapses. Visual–audio consistency remains high for both. Global consistency is, on average, higher for real videos, but the asymmetry is the distinguishing factor.

Underlying Causes of Asymmetric Cross-Modal Consistency This asymmetry is consistent with the production mechanics of "cheapfakes" (Paris and Donovan, 2019). Qualitatively, we observe that fake news producers often start with high-

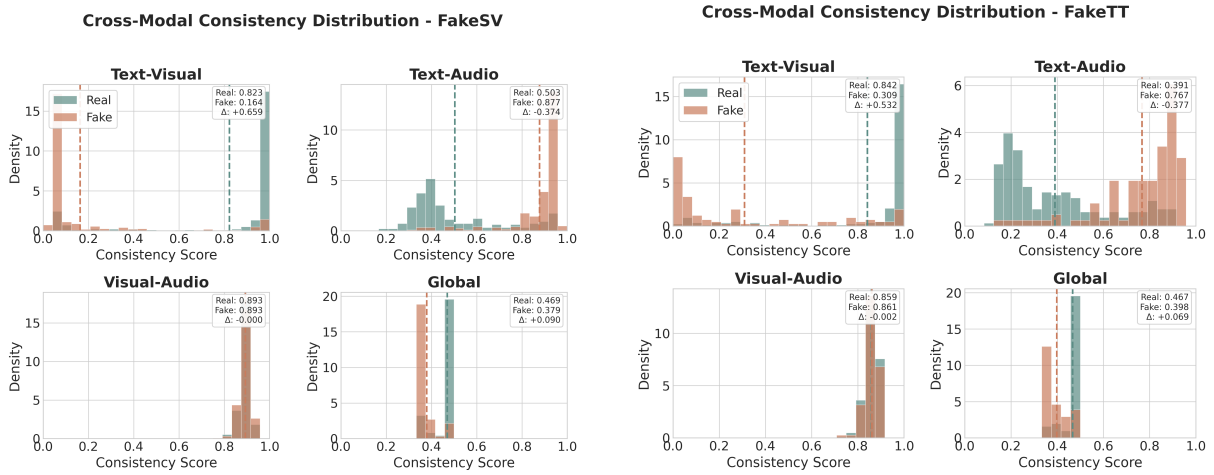

Figure 3: Cross-modal consistency distributions of text–visual, text–audio, visual–audio, and global consistency scores for real vs. fake videos. Real news shows high text–visual but moderate text–audio consistency; fake news flips this pattern, while visual–audio consistency remains high. Left: FakeSV; right: FakeTT.

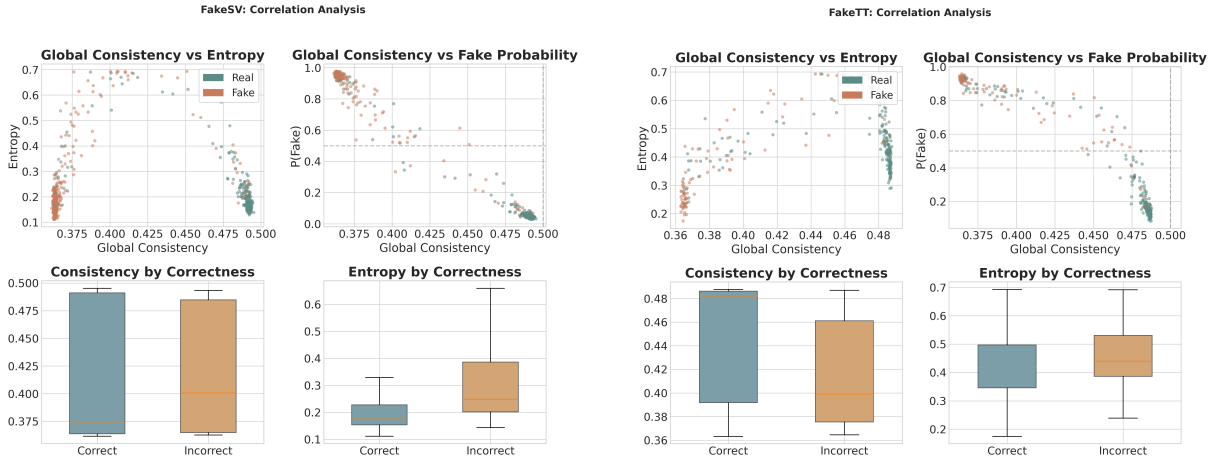

Figure 4: Consistency–prediction analysis. Global consistency is strongly (negatively) correlated with fake probability; errors concentrate in the middle-consistency band, while c_{global} vs. entropy is non-linear with uncertainty peaking at mid-range consistency. Left: FakeSV; right: FakeTT.

quality, visually coherent stock footage or stolen clips, which naturally preserves strong alignment between the visual stream and background audio. These materials are then repurposed with fabricated subtitles and voiceovers to convey a false narrative. The generated text is designed to be emotionally engaging and closely aligned with the affect of the audio, while gradually drifting away from the actual visual evidence, creating the semantic gap that MAGIC^3 detects. In contrast, real news videos, particularly those from professional outlets in FakeTT, emphasize factual grounding, with voiceovers that closely describe the visual events, resulting in stronger text–visual consistency. More statistics are shown in Appendix D.1.

5.2 Global Consistency for Interpretability

We examine how the global consistency score c_{global} relates to predictions. Figure 4 plots c_{global} against the predicted fake probability on FakeSV (left) and FakeTT (right). c_{global} is strongly and roughly *negatively* correlated with the model's fake probability; prediction errors cluster in the middle band of c_{global}, indicating that the score provides an interpretable, low-dimensional difficulty axis.

Interpretability of Global Consistency Because c_{global} is an explicit input feature, this correlation is expected. However, its value lies in its *operational utility*. Unlike opaque end-to-end probabilities, c_{global} offers human moderators a tangible explanation for the model's decision: a video is flagged not just because it "looks fake," but specif-

	Variant	Acc	F1	AUC
FakeSV				
	MAGIC^3 (full)	**86.71**	**86.84**	92.23
Consistency	w/o CMCG	84.90	84.85	91.77
	w/o CFE	85.69	85.61	92.30
	w/o TCMI	85.75	85.89	91.94
Fusion	w/o AARF	84.49	84.71	91.43
	w/o HMT	85.86	85.79	**92.48**
Training	w/o contr.	85.75	85.89	92.15
	w/o adv.	86.56	86.53	92.23
FakeTT				
	MAGIC^3 (full)	**84.95**	**82.28**	**89.08**
Consistency	w/o CMCG	81.94	78.68	87.44
	w/o CFE	78.83	80.13	87.88
	w/o TCMI	78.64	77.22	86.92
Fusion	w/o AARF	80.91	78.90	88.30
	w/o HMT	82.91	81.48	88.47
Training	w/o contr.	82.93	80.47	88.50
	w/o adv.	78.37	77.78	88.37

Table 2: Ablation study. Consistency modules (CMCG, CFE, TCMI) and AARF are critical.

ically because its modalities fail to cohere. We show that uncertainty peaks precisely in the middle-consistency region in Appendix D.2, suggesting that c_{global} can serve as a filter to identify ambiguous cases that require human or VLM intervention.

5.3 Uncertainty-Aware Two-Stage VLM Routing

We now examine how consistency and uncertainty can be used to orchestrate a two-stage pipeline with a heavyweight VLM detector. Figure 5 summarizes uncertainty distributions, calibration curves, and routing behavior on both datasets. We implement FakeSV-VLM (Wang et al., 2025c) for the stage-2 detector. The top row shows that entropy is typically lower for correct predictions than for incorrect ones, and that confidence is reasonably calibrated. The bottom row then visualizes routing policies that identify difficult samples for VLM escalation. We treat the routing threshold as a hyperparameter and tune it on the validation set to route approximately 25% of samples on both datasets.

Routed samples are indeed harder (lower detector-only accuracy), and the overall accuracy of the two-stage system improves. As shown in Table 1, on FakeSV, routing 25% of samples to a VLM (MAGIC^3+VLM) increases overall accuracy from 86.71% to 90.93%, surpassing FakeSV-VLM's 89.60% by 1.33 percentage points. On FakeTT, a 25% routing ratio boosts accuracy from 84.95% to 89.52%, surpassing the VLM-only baseline of 88.90% by 0.62pp. Crucially, this hybrid approach achieves 18–27$\times$ higher throughput and massive VRAM savings compared to a pure VLM pipeline. MAGIC^3 effectively filters out the bulk of easy cases ($\sim$75%), allowing expensive VLMs to focus only on the hardest samples, and thus achieving the best cost–performance tradeoff.

5.4 Consistency Stability under Style Perturbations

Finally, we explore the effect of multi-style LLM rewrites. Adversarial-Aware Rewrite Fusion fuses the original caption with neutral, formal, and sensational rewrites and is trained with style-aware contrastive losses. Table 2 reports an ablation in which we remove AARF (i.e., use only the original text) while keeping other components fixed. Performance drops on both datasets, indicating that exposure to style variation improves robustness. Conceptually, this connects to work on style-aware neural fake news detection like Style-News (Wang et al., 2024a), where stylized generation is used as an adversary to stress-test detectors. Our setup differs in two ways: we treat rewrites as semantic-preserving perturbations generated by a general-purpose LLM (rather than a task-specific generator), and we explicitly measure how cross-modal consistency behaves under style changes (statistical details in Appendix D.3). Our hypothesis is that real news remains consistent under style rewrites, whereas fake news shows larger variance, especially in text–visual consistency. This is supported qualitatively by Figure 7, where fake samples exhibit larger shifts under stylistic perturbations. Appendix D.3 reports detailed variance statistics.

6 Conclusion

We presented MAGIC^3, a consistency-centric detector for fake news short videos that explicitly exposes cross-tri-modal consistency at multiple granularities: pairwise and global scores, token-level consistency fields, a temporal audio–visual inconsistency score, and uncertainty estimates. On FakeSV and FakeTT, MAGIC^3 achieves state-of-the-art supervised performance on frozen features and offers a superior cost–performance tradeoff: when combined with VLM routing, it surpasses VLM-only accuracy at 18–27$\times$ higher throughput. Future work should integrate external knowledge search for fact-level verification.

Limitations

First, our method relies on pre-extracted multimodal features and offline LLM-generated text rewrites, which introduce additional pre-processing cost and potential sensitivity to the choice of LLMs and prompt designs; however, this is a deliberate design choice that decouples heavy backbone computation from the detector head, enabling high throughput, low memory usage, and flexible integration with different pretrained encoders, and our ablation results further show that the core consistency modeling remains effective even without LLM rewrites. Second, the current TCMI module focuses on coarse-grained temporal audio–visual inconsistency and does not explicitly model fine-grained temporal logic or causal structures; nevertheless, it is intentionally designed as a lightweight complement to static consistency signals, targeting prominent synchronization mismatches commonly observed in repurposed or cheaply manipulated videos. Third, MAGIC^3 operates as a feature-level detector and does not directly process raw video inputs; while this limits strictly on-device, end-to-end deployment, it allows the framework to remain backbone-agnostic and deployment-friendly for large-scale platform settings where features are often cached or shared across tasks. Moreover, consistency-based detectors may still struggle with carefully crafted misinformation in which all modalities are internally consistent but factually incorrect; addressing such cases typically requires external knowledge and verification resources, and we view our consistency-centric approach as a complementary detection axis rather than a standalone solution to factual verification. Finally, although the proposed consistency fields and uncertainty estimates improve model transparency and enable fine-grained analysis, more human-centered evaluations are needed to fully understand how practitioners interpret and act upon these signals, which we leave as an important direction for future work.

Ethics Statement

This work aims to mitigate the harmful impact of multimedia misinformation in short-form news videos. Nonetheless, there are dual-use risks: understanding detection mechanisms can inform adversaries seeking to evade them. We therefore focus on general architectural principles rather than attack-specific heuristics, and we encourage responsible deployment practices that combine automated detection with human oversight.

All datasets used are publicly available; we follow their original licenses and respect existing privacy safeguards. In real-world deployments, particular care should be taken to monitor false positives, ensure fairness across topics and languages, and provide appropriate channels for appeals when automated decisions affect users.

References

Shahin Amiriparian, Filip Packań, Maurice Gerczuk, and Björn W Schuller. 2024. Exhubert: Enhancing hubert through block extension and fine-tuning on 37 emotion datasets. *arXiv preprint arXiv:2406.10275*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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*, pages 8770–8780.

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. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Hyewon Choi and Youngjoong Ko. 2021. Using topic modeling and adversarial neural networks for fake news video detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2950–2954.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yuan Gao, Xuelong Wang, Yu Zhang, Ping Zeng, and Yingjie Ma. 2024. Temporal feature prediction in audio–visual deepfake detection. *Electronics*, 13(17):3433.

Xuejian Huang, Tinghuai Ma, Hao Tang, and Huan Rong. 2025. Knowledge-enhanced dynamic scene

graph attention network for fake news video detection. *IEEE Transactions on Multimedia*.

Xianghua Li, Jiao Qiao, Shu Yin, Lianwei Wu, Chao Gao, Zhen Wang, and Xuelong Li. 2025. A survey of multimodal fake news detection: a cross-modal interaction perspective. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 16–25.

Britt Paris and Joan Donovan. 2019. Deepfakes and cheap fakes. *United States of America: Data & Society*, 1.

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

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE international conference on big data (big data)*, pages 899–908. IEEE.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320.

Chong Tian, Qirong Ho, and Xiuying Chen. 2025. A symbolic adversarial learning framework for evolving fake news generation and detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12307–12321.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Haoran Wang, Yan Yang, and Yingli Zhong. 2025a. Global and local feature enhancement for short video fake news detection. In *International Conference on Intelligent Computing*, pages 435–446. Springer.

Junxi Wang, Jize Liu, Na Zhang, and Yaxiong Wang. 2025b. Consistency-aware fake videos detection on short video platforms. In *International Conference on Intelligent Computing*, pages 200–210. Springer.

Junxi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. 2025c. Fakesv-vlm: Taming vlm for detecting fake short-video news via progressive mixture-ofexperts adapter. *arXiv preprint arXiv:2508.19639*.

Wei-Yao Wang, Yu-Chieh Chang, and Wen-Chih Peng. 2024a. Style-news: Incorporating stylized news generation and adversarial verification for neural fake news detection. *arXiv preprint arXiv:2401.15509*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and 1 others. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

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. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

Yukun Yang, Xiwei Shi, Haoxu Li, Buwei Fan, and Yijia Xu. 2025. Fake news detection in short videos by integrating semantic credibility and multi-granularity contrastive learning. *Applied Sciences*, 15(23):12621.

X Zhou, J Wu, and R Zafarani. 2020. Safe: similarity-aware multi-modal fake news detection. arxiv. *arXiv preprint arXiv:2003.04981*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A Cross-modal Inconsistency Taxonomy (Details)

Table 3 provides a detailed version of the task-oriented inconsistency taxonomy mentioned in the introduction. For each category we list a brief description and typical patterns; MAGIC^3's corresponding signals are discussed in the main text.

B Additional Figures

See Figure 5 and Figure 7.

C Method Details

This section provides the mathematical details of the components sketched in Section 3.

C.1 Cross-Modal Consistency Gate (CMCG)

We first compute pooled representations for each modality:

$$\bar{\mathbf{h}}_{\text{text}} = \text{Pool}(\mathbf{H}_{\text{text}}), \quad \bar{\mathbf{h}}_{\text{vis}} = \text{Pool}(\mathbf{H}_{\text{vis}}), \\ \bar{\mathbf{h}}_{\text{aud}} = \text{Pool}(\mathbf{H}_{\text{aud}}). \tag{11}$$

Here $\text{Pool}(\cdot)$ denotes mean pooling or a special CLS token.

For each modality pair $(m_i, m_j) \in \{(\text{text}, \text{vis}), (\text{text}, \text{aud}), (\text{vis}, \text{aud})\}$, CMCG computes a scalar consistency score

$$c_{ij} = \sigma\Big(\mathbf{w}_{ij}^\top [\bar{\mathbf{h}}_{m_i}; \bar{\mathbf{h}}_{m_j}] + b_{ij}\Big), \tag{12}$$

where $[\cdot;\cdot]$ is concatenation and σ is the sigmoid function. This yields a 3-dimensional consistency vector

$$\mathbf{c} = [c_{\text{tv}}, c_{\text{ta}}, c_{\text{va}}] \in \mathbb{R}^3. \tag{13}$$

A small MLP then produces a global consistency score

$$c_{\text{global}} = \sigma(\text{MLP}_c(\mathbf{c})). \tag{14}$$

C.2 Consistency Field Estimator (CFE)

Suppose we have a cross-attention matrix from modality m_i to m_j,

$$\mathbf{A}_{m_i \leftarrow m_j} \in \mathbb{R}^{L_{m_i} \times L_{m_j}}, \tag{15}$$

produced by the cross-modal transformer. Row t corresponds to a token or frame in modality m_i. We define the consistency field of m_i with respect to m_j as

$$F_{ij}^{(m_i)}(t) = \max_k \mathbf{A}_{m_i \leftarrow m_j}(t, k), \quad t = 1, \ldots, L_{m_i}. \tag{16}$$

Intuitively, $F_{ij}^{(m_i)}(t) \in [0, 1]$ reflects how strongly token t in modality m_i finds supporting evidence in modality m_j. For training, we add a mild regularizer aligning global CMCG scores with aggregated fields:

$$\mathcal{L}_{\text{reg}} = \sum_{(i,j)} \big(c_{ij} - \text{mean}(F_{ij}^{(m_i)})\big)^2. \tag{17}$$

C.3 Multi-view Adversarial-Aware Rewrite Fusion (AARF)

Let $\mathbf{H}_{\text{text}}$ be the original text sequence and $\mathbf{H}_{\text{rew}}^{(v)}$ the v-th rewrite. We pool them into vectors

$$\mathbf{h}_{\text{orig}} = \text{Pool}(\mathbf{H}_{\text{text}}), \tag{18}$$

$$\mathbf{h}_{\text{rew}}^{(v)} = \text{Pool}\big(\mathbf{H}_{\text{rew}}^{(v)}\big). \tag{19}$$

For each view v, we estimate a scalar rewrite quality

$$q^{(v)} = \sigma\big(\text{MLP}_q[\mathbf{h}_{\text{orig}}; \mathbf{h}_{\text{rew}}^{(v)}]\big), \tag{20}$$

and compute gating logits

$$\boldsymbol{g} = \text{MLP}_g\big([\mathbf{h}_{\text{orig}}; \mathbf{h}_{\text{rew}}^{(1)}; \ldots; \mathbf{h}_{\text{rew}}^{(V)}; q^{(1)}, \ldots, q^{(V)}]\big). \tag{21}$$

We obtain preliminary weights $\boldsymbol{\alpha}^{\text{soft}} = \text{softmax}(\boldsymbol{g})$, enforce a minimum weight $\alpha_{\min}$ on the original text, and renormalize:

$$\tilde{\alpha}_0 = \max(\alpha_0^{\text{soft}}, \alpha_{\min}), \tag{22}$$

$$\tilde{\alpha}_v = \alpha_v^{\text{soft}}, \quad v \geq 1, \tag{23}$$

$$\alpha_v = \begin{cases} \tilde{\alpha}_0, & v = 0, \\ (1 - \tilde{\alpha}_0)\, \dfrac{\tilde{\alpha}_v}{\sum_{u=1}^{V} \tilde{\alpha}_u}, & v \geq 1. \end{cases} \tag{24}$$

The fused representation is

$$\mathbf{h}_{\text{text}}^{\text{fuse}} = \alpha_0\, \mathbf{h}_{\text{orig}} + \sum_{v=1}^{V} \alpha_v\, \text{Proj}\big(\mathbf{h}_{\text{rew}}^{(v)}\big). \tag{25}$$

An InfoNCE-style loss encourages all views of the same video to be close and views of different videos to be separated.

C.4 Hierarchical Multimodal Transformer (HMT)

Layer A performs intra-modal refinement:

$$\mathbf{H}_{\text{text}}^{(A)} = \text{TransEnc}_{\text{text}}(\mathbf{H}_{\text{text}}), \tag{26}$$

$$\mathbf{H}_{\text{vis}}^{(A)} = \text{TransEnc}_{\text{vis}}(\mathbf{H}_{\text{vis}}), \tag{27}$$

$$\mathbf{H}_{\text{aud}}^{(A)} = \text{TransEnc}_{\text{aud}}(\mathbf{H}_{\text{aud}}). \tag{28}$$

Category	Description and typical patterns
Entity & context inconsistency	Entities (persons, organizations, places) or contexts (e.g., country, city, venue) mentioned in text do not match those present in video or audio. A typical pattern is subtitles claiming "thousands protest in Paris" while the video shows a small rally in a different city.
Event-level semantic inconsistency	Modalities disagree on the core event: what happened, at what scale, or with which participants. For example, the caption claims "violent clashes erupt" but the video shows a peaceful gathering, or the audio describes a different incident.
Affective inconsistency	Sentiment or emotion conveyed by audio contradicts text or visual content, such as tragic subtitles overlaid on upbeat background music or calm narration over highly dramatic visuals.
Temporal audio–visual inconsistency	Audio and visual streams are misaligned in time, reused from different footage, or re-ordered. This includes dubbed videos, re-edited footage, or mis-timed sound effects.
Stylistic / pragmatic inconsistency	Rhetorical style, framing, or level of sensationalism in one modality is inconsistent with the others, e.g., a sensational subtitle for a relatively mundane video, or highly emotional text with neutral audio.

Table 3: Task-oriented taxonomy of inconsistencies in short-form news videos. These categories serve as design targets for MAGIC^3's consistency-centric architecture.

Layer B performs consistency-weighted cross-modal attention; for modalities m_i, m_j we compute

$$\tilde{\mathbf{H}}_{m_i \leftarrow m_j} = \mathrm{CrossAttn}\big(\mathbf{H}^{(A)}_{m_i}, \mathbf{H}^{(A)}_{m_j}\big). \tag{29}$$

Attention weights are stored for CFE. We then aggregate messages using weights derived from $\mathbf{c}$:

$$\boldsymbol{\beta}_m = \mathrm{softmax}\big(\mathrm{MLP}_m(\mathbf{c})\big), \tag{30}$$

$$\mathbf{H}^{(B)}_m = \sum_{j \neq m} \beta_{mj}\, \tilde{\mathbf{H}}_{m \leftarrow j}. \tag{31}$$

Layer C pools refined representations and fuses them with the AARF text:

$$\mathbf{Z} = \begin{bmatrix} \mathrm{Pool}(\mathbf{H}^{(B)}_{\text{text}}) \\ \mathrm{Pool}(\mathbf{H}^{(B)}_{\text{vis}}) \\ \mathrm{Pool}(\mathbf{H}^{(B)}_{\text{aud}}) \\ \mathbf{h}^{\text{fuse}}_{\text{text}} \end{bmatrix}, \tag{32}$$

$$\mathbf{h}_{\text{global}} = \mathrm{TransEnc}_{\text{global}}(\mathbf{Z})_{[\text{CLS}]}. \tag{33}$$

C.5 Contrastive–Adversarial Joint Learning (CAJL)

Let $y_i \in \{0, 1\}$ be labels and $\hat{y}_i$ predicted probabilities. The cross-entropy loss is

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N} \sum_i \Big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big). \tag{34}$$

We use InfoNCE for contrastive terms. For representation pairs $(\mathbf{u}_i, \mathbf{v}_i)$ from the same video and negatives from other videos,

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N} \sum_i \log \frac{\exp(\mathrm{sim}(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\sum_j \exp(\mathrm{sim}(\mathbf{u}_i, \mathbf{v}_j)/\tau)}. \tag{35}$$

$\mathcal{L}_{\text{intra}}$ uses different views of the same video (original text, rewrites, audio, visuals), and $\mathcal{L}_{\text{cross}}$ aligns fused text and global representations.

For adversarial consistency, we apply small perturbations δ_i to intermediate representations and enforce prediction consistency:

$$\mathcal{L}_{\text{adv}} = \frac{1}{N} \sum_i \mathrm{KL}\big(p_\theta(y \mid \mathbf{h}_i) \| p_\theta(y \mid \mathbf{h}_i + \delta_i)\big). \tag{36}$$

We also encourage consistent predictions between original and fused text:

$$\mathcal{L}_{\text{sem}} = \frac{1}{N} \sum_i \mathrm{KL}\big(p_\theta(y \mid \mathbf{h}_{\text{orig},i}) \| p_\theta(y \mid \mathbf{h}^{\text{fuse}}_{\text{text},i})\big). \tag{37}$$

A style invariance term applies InfoNCE over projections of all text views:

$$\mathcal{L}_{\text{style}} = \frac{1}{N} \sum_i \mathcal{L}_{\text{NCE}}\big(\{\mathbf{z}^{(v)}_{\text{text},i}\}_{v=0}^{V}\big). \tag{38}$$

The final loss is

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{39}$$

Hyperparameters are given in Table 11.

D Additional Quantitative Analyses for Findings

This appendix reports the quantitative statistics that support Findings 1–4.

D.1 Aggregate Consistency Statistics

Table 5 lists mean ± standard deviation of the four consistency scores for real vs. fake videos,

Component	Function	Consistency level	Role in findings / interpretability
Consistency Computation			
CMCG	Computes pairwise consistency scores $[c_{\text{tv}}, c_{\text{ta}}, c_{\text{va}}]$ and global score c_{global}	Pairwise → global	Reveals asymmetric text–visual vs. text–audio patterns; defines the global consistency axis (Figures 3–4).
CFE	Derives token/frame-level consistency fields from cross-modal attention	Local (token / frame)	Localizes suspicious phrases or frames; supports heatmap visualization (Figure 6).
TCMI	Computes temporal audio–visual inconsistency score c_{temp}	Temporal	Detects coarse audio–visual synchronization mismatches common in repurposed videos.
Cross-Modal Fusion			
AARF	Fuses original text with neutral/formal/sensational LLM rewrites via gated weights	Text-internal	Builds style-robust text embeddings; used to study style perturbation effects (Finding 4).
HMT	Three-stage hierarchical fusion: intra-modal refinement, consistency-weighted cross-attention, global aggregation	Structural	Integrates multimodal representations with consistency-modulated attention.
Training and Prediction			
CAJL	Contrastive–adversarial joint learning with InfoNCE and perturbation losses	Training	Aligns multimodal representations and improves robustness under perturbations.
Classifier	Outputs fake probability, confidence, and uncertainty for two-stage routing	Prediction	Provides calibrated uncertainty for selective VLM escalation (Finding 3).

Table 4: MAGIC^3 component overview. Components are grouped by function: consistency computation (CMCG, CFE, TCMI), cross-modal fusion (AARF, HMT), and training/prediction (CAJL, Classifier). Together they provide the signals used in our four findings and the two-stage VLM pipeline.

along with two-sample t-tests. The text–visual vs. text–audio asymmetry is highly significant on both datasets; visual–audio is nearly identical across classes.

Dataset	Score	Real	Fake	p-value
FakeSV	c_{tv}	0.82 ± 0.32	0.16 ± 0.27	$p \ll 0.001$
	c_{ta}	0.50 ± 0.20	0.88 ± 0.13	$p \ll 0.001$
	c_{va}	0.89 ± 0.02	0.89 ± 0.02	$p = 0.82$
	c_{global}	0.47 ± 0.04	0.38 ± 0.04	$p \ll 0.001$
FakeTT	c_{tv}	0.84 ± 0.29	0.31 ± 0.35	$p \ll 0.001$
	c_{ta}	0.39 ± 0.23	0.77 ± 0.18	$p \ll 0.001$
	c_{va}	0.86 ± 0.03	0.86 ± 0.03	$p = 0.49$
	c_{global}	0.47 ± 0.04	0.40 ± 0.04	$p \ll 0.001$

Table 5: Finding 1 statistics: consistency scores by class (mean±std) and two-sample t-test p-values. Text–visual vs. text–audio patterns flip between real and fake videos on both datasets.

D.2 Correlation Between Global Consistency and Predictions

Table 6 reports Spearman/Pearson correlations between c_{global} and predicted fake probability, along with a linear-fit R^2. Table 7 shows the empirical fake rate in five c_{global} bins, illustrating the generally decreasing trend.

Dataset	Spear. ρ	Pear. r	R^2
FakeSV	-0.91 ($p \ll 0.001$)	-0.99 ($p \ll 0.001$)	0.99
FakeTT	-0.96 ($p \ll 0.001$)	-0.96 ($p \ll 0.001$)	0.91

Table 6: Finding 2: correlation between c_{global} and predicted fake probability.

c_{global} bin	FakeSV rate	FakeTT rate
Lowest (Q1)	0.93	0.83
Q2	0.89	0.55
Q3	0.76	0.17
Q4	0.19	0.03
Highest (Q5)	0.05	0.07

Table 7: Finding 2: fake rates in five quantile bins of c_{global} (lower scores = more inconsistent).

D.3 Style Sensitivity Statistics

Table 8 compares style-induced variance (across original+three rewrites) for real vs. fake videos. Fake samples show consistently higher variance for text–visual, text–audio, and global consistency on both datasets; visual–audio is unchanged, as expected.

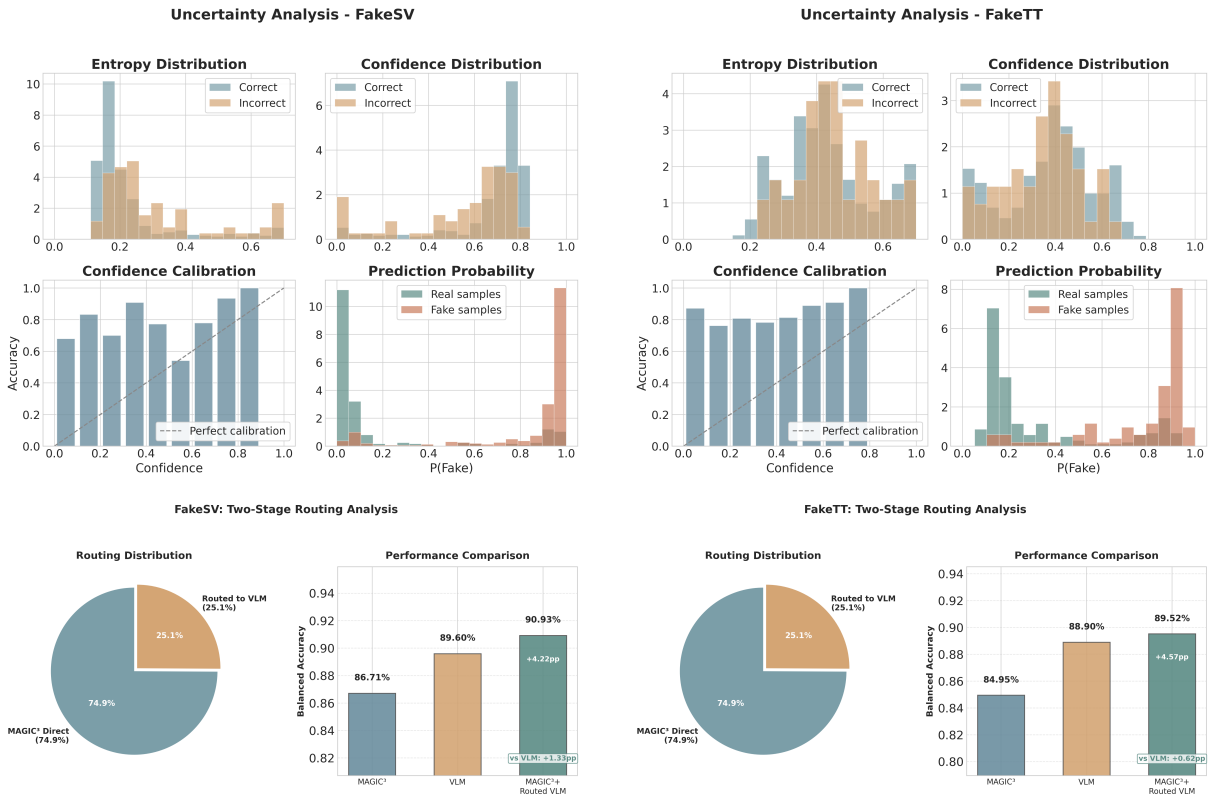

Figure 5: Uncertainty and two-stage routing behaviour of MAGIC^3. Top: entropy/confidence distributions and calibration plots. Bottom: split between direct predictions and routed samples, and accuracies for routed vs. non-routed subsets when using a VLM-based stage-2 detector. Left: FakeSV; right: FakeTT.

Dataset	Score	Real var	Fake var	p
FakeSV	c_{tv}	0.0336	0.0815	$p < 10^{-16}$
	c_{ta}	0.0109	0.0148	$p < 10^{-2}$
	c_{va}	0.0	0.0	n.s.
	c_{global}	6.19×10^{-4}	1.44×10^{-3}	$p < 10^{-15}$
FakeTT	c_{tv}	0.0194	0.0801	$p < 10^{-21}$
	c_{ta}	0.0081	0.0170	$p < 10^{-7}$
	c_{va}	0.0	0.0	n.s.
	c_{global}	3.13×10^{-4}	1.23×10^{-3}	$p < 10^{-20}$

Table 8: Finding 4: variance of consistency scores under style rewrites (real vs. fake). "n.s." denotes non-significant differences ($p \geq 0.05$).

D.4 Two-Stage Routing Analysis Details

We treat the routing threshold as a hyperparameter and tune it on the validation set to route approximately 25% of samples on both datasets. Table 9 summarizes the routing strategies used. For FakeSV, we use entropy-based routing (high entropy indicates uncertain predictions); for FakeTT, we use a *difficulty score* that combines entropy, global consistency, and confidence: $d = \text{entropy}_{\text{norm}} + (1 - c_{\text{global,norm}}) + (1 - \text{conf}_{\text{norm}})$, routing samples with the highest difficulty scores.

Dataset	Routed Ratio	MAGIC^3+VLM Acc	Routing Strategy
FakeSV	25.1%	90.93%	Entropy-based
FakeTT	25.1%	89.52%	Difficulty score

Table 9: Finding 3: routing strategies and routed ratios used to construct the two-stage systems. Both strategies outperform the standalone VLM detector (FakeSV: +1.33pp, FakeTT: +0.62pp).

E Datasets and Splits

Table 10 summarizes the two benchmarks used in this paper. We follow the official chronological splits from Huang et al. (2025), using 70%/15%/15% for training, validation, and testing.

Statistic	**FakeSV**	**FakeTT**
Time range	2017/10–2022/02	2019/05–2024/03
Avg. duration (s)	39.88	47.69
# Fake videos	1,810	1,172
# Real videos	1,814	819
# Total	3,624	1,991

Table 10: Statistics of FakeSV and FakeTT. All videos are short-form clips on real-world platforms.

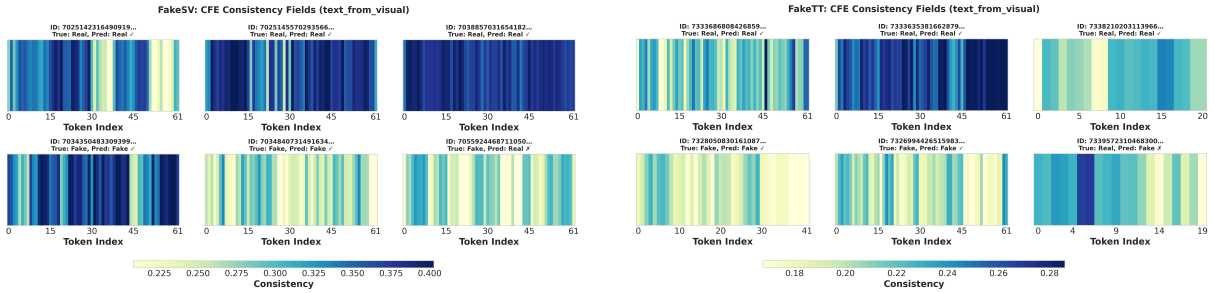

Figure 6: CFE consistency fields. Colors indicate per-token consistency with the visual stream; sharp high-score regions suggest localized support, while diffuse/low scores indicate potential inconsistencies.

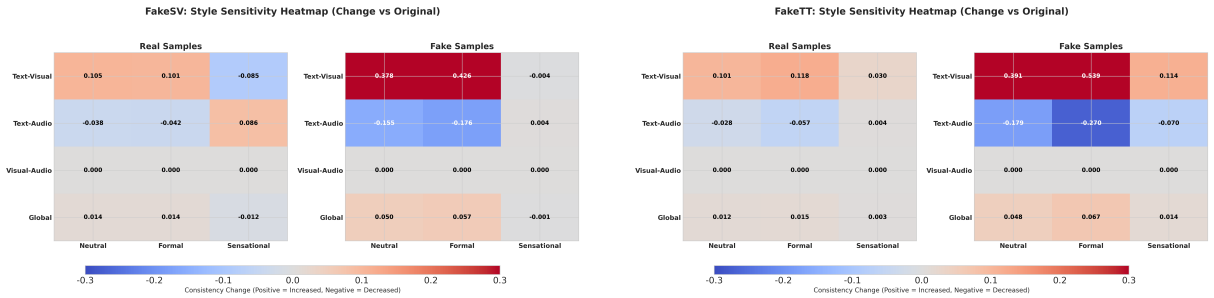

Figure 7: Style sensitivity heatmaps (Finding 4). Rows: consistency types; columns: style rewrites vs. original. Fake samples show larger shifts than real ones.

F LLM Rewriting Prompts

We use three prompts to obtain multi-view rewrites. In our experiments we use DeepSeek-V3.2; for FakeSV we use Chinese translations of these prompts.

Neutral news style.

> You are a professional news editor.
> Rewrite the following video description in a neutral, objective news style.
> Focus on clarity and factual accuracy. Do not add new information.
>
> Original text: {input_text}
>
> Rewritten text:

Formal elaborated style.

> You are a senior journalist.
> Rewrite the following video description in a more formal and elaborated style,
> making the key facts explicit and clearly structured. Avoid emotional language.
>
> Original text: {input_text}
>
> Rewritten text:

Sensational style.

> You are writing a sensational social media post about breaking news.
> Rewrite the following video description to sound more emotional and attention-grabbing,
> while keeping the core facts unchanged.
>
> Original text: {input_text}
>
> Rewritten text:

G Implementation Details and Hyperparameters

Backbone encoders use `bert-base-multilingual-uncased` for text (maximum length 256), Swin Transformer (Liu et al., 2021) for K keyframes per video, and ExHuBERT-based emotion embeddings for audio (Amiriparian et al., 2024). These encoders are used only for offline feature extraction; during MAGIC^3 training we operate purely on the pre-extracted features and do not update encoder parameters.

Training uses AdamW with learning rate 5×10^{-5}, batch size 32, cosine decay with 10% warm-up, dropout 0.3 in transformer layers, and early stopping based on validation macro-F1. We train for up to 50 epochs and keep the checkpoint with best validation macro-F1. In our main experiments we set $\lambda_{\text{intra}} = \lambda_{\text{cross}} = 0.1$, $\lambda_{\text{adv}} = 0.5$, $\lambda_{\text{sem}} = 0.1$, $\lambda_{\text{style}} = 0.1$, and $\lambda_{\text{reg}} = 0.05$, with $\alpha_{\min} = 0.5$, $V = 3$, and $K_{\text{MC}} = 5$.

H Efficiency and Deployment Discussion

Because all encoders are used only for offline feature extraction and MAGIC^3 operates on pre-extracted features, inference on cached features reduces to a single detector head. The trainable

Hyperparameter	Value
Learning rate	5×10^{-5}
Batch size	32
Epochs	50
Warmup ratio	0.1
Hidden dimension	256
Attention heads (HMT)	8
Dropout	0.3
λ_{intra}	0.1
λ_{cross}	0.1
λ_{adv}	0.5
λ_{sem}	0.1
λ_{style}	0.1
λ_{reg}	0.05
$\alpha_{\min}$ (AARF)	0.5
V (rewrites per video)	3

Table 11: Key hyperparameter settings for MAGIC^3.

detector head has roughly 185M parameters (with no encoder fine-tuning), which is lightweight compared to multi-billion-parameter VLM backbones. On a single RTX 5090 GPU, a full training run on each dataset (including validation and test evaluation) takes about 15 minutes. With inference batch size 128 on the same GPU, we measure about 680 samples/s and $\sim$1.5 ms per sample, with $\sim$1.67 GB peak GPU memory on cached features. Peak GPU memory usage during training is about 11 GB on the same GPU. End-to-end latency is dominated by video decoding and feature extraction, which can be amortized via caching and asynchronous preprocessing in practical systems.

In contrast, VLM-based detectors typically require end-to-end processing with a large vision–language backbone, which is more expensive to train (often requiring multi-GPU setups) and to serve at scale. Operating purely at the feature level allows MAGIC^3 to be retrained quickly for new topics or sub-verticals within the short-video misinformation space, making it a practical choice for platforms that need rapid adaptation under hardware constraints.