From What Is Said to Why It Is Framed: Intent-Aware News Video Understanding

Xiangzheng Kong[1,2] **Minnan Luo**[1,2*]
Wenya Wang[3] **Jiaying Wu**[4] **Zhi Zeng**[1,2] **Guang Dai**[5]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3]Nanyang Technological University, Singapore
[4]National University of Singapore, Singapore
[5]SGIT AI Lab, State Grid Corporation of China

Abstract

Short-form news videos increasingly shape public perception through strategic framing, yet existing verification methods largely overlook the *communicative intent* underlying such content. By emphasizing surface semantics, current models struggle to separate stylistic presentation from factual evidence, which leads to shortcut learning and brittle generalization. To address this limitation, we propose the Origin–Objective–Means (OOM) framework, a theory-grounded representation of communicative intent that captures creator stance, audience need activation, and communication strategy. We validate OOM through large-scale human annotation, revealing distinct and consistent lexical and structural patterns across intent dimensions. Building on this representation, we operationalize intent as an explicit semantic condition rather than a prediction target. Concretely, we introduce Intent-Guided Prompting (IGP) to condition LLM reasoning and intent-conditioned multimodal detection framework **(ICMD)**, which injects intent into multimodal detectors via feature-wise modulation. Experiments on FakeSV and FakeTT show that modeling intent as an intermediate condition consistently improves accuracy and robustness across diverse vision–language backbones, while substantially reducing reliance on spurious stylistic correlations.

1 Introduction

Short-form news videos have become a dominant medium for public information dissemination. Unlike traditional text-based reporting, video news combines visual framing, audio cues, narrative pacing, and affective signals, enabling strategic presentation beyond factual reporting (Kang, 2019). As a result, videos convey not only what is reported but also **why it is framed in a particular way** (Jo et al., 2024; Yerukola et al., 2024; Edstedt et al., 2022; Ha et al., 2024; Chen et al., 2022; Qian et al., 2021; Hong et al., 2025; Zhang et al., 2025c; Zeng et al., 2025c,b; Ma et al., 2025; Wan et al., 2025). This capacity for strategic framing complicates automated verification, as stylistic presentation and factual manipulation are often intertwined.

Figure 1: Overview of our Origin–Objective–Means (OOM) framework, which models news video intent as an explicit intermediate semantic representation.

Despite progress in multimodal news understanding, most existing approaches focus on surface-level semantics, such as topic classification, sentiment analysis, or direct veracity prediction (Mittal et al., 2024; Xiao et al., 2024; Weld et al., 2022; Gabriel et al., 2022; Zeng et al., 2026; Tong et al., 2025; Xu et al., 2025). While effective at capturing content meaning, these methods largely overlook communicative intent. Consequently, models often conflate benign stylistic choices with manipulative tactics, relying on spurious correlations between production style and veracity rather than reasoning over factual evidence. This shortcut learning degrades robustness and interpretability, especially in emotionally charged news videos.

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2026, pages 39039–39050

We argue that addressing these limitations requires explicit modeling of **communicative intent**. In practice, news production is a goal-driven communicative act in which creators adopt specific stances and employ strategies to activate audience needs (Ecker et al., 2022). Without a structured representation of this intent layer, multimodal systems lack a principled mechanism to explain why particular visual or auditory cues are emphasized or amplified.

To this end, we propose the **Origin–Objective–Means (OOM)** framework, a theory-grounded representation of news video intent (Figure 1). Drawing on evolutionary communication and motivational psychology (Maslow, 1943; Silvey, 2016; Sperber and Wilson, 1986), OOM decomposes communicative intent into three complementary dimensions. **(1) Origin** captures the creator's stance toward an event, such as supportive, neutral, or critical, defining the perspective from which the narrative is constructed. This stance is instantiated through an **(2) Objective**, which models the psychological needs the content aims to activate, ranging from survival and safety concerns to higher-level motivations such as social belonging, esteem, and cognitive understanding. These objectives are realized through **(3) Means**, which describe the rhetorical strategies used to frame the content, including factual presentation, ideological framing, and emotional mobilization. Together, these dimensions form a minimal yet expressive semantic abstraction that captures the communicative goals of news videos beyond surface-level features.

Notably, we do not treat intent as another classification target. Instead, we operationalize OOM as an explicit intermediate semantic condition. This design allows intent to function as a computational interface that guides how multimodal evidence is interpreted, rather than serving as a direct predictor of veracity. To instantiate this idea, we introduce **Intent-Guided Prompting (IGP)**, which structures Large Language Model (LLM) reasoning through the OOM lens, as well as **ICMD**, an intent-conditioned multimodal detection framework that injects intent signals into multimodal detection networks via feature-wise modulation.

We instantiate the OOM framework through large-scale multi-annotator labeling on FakeSV (Qi et al., 2023) and FakeTT (Bu et al., 2024), two real-world video misinformation datasets. Our analysis shows that OOM categories exhibit distinct lexical and structural patterns, confirming that the framework is grounded in authentic communicative practices. Experiments across diverse vision–language backbones demonstrate that intent-conditioned reasoning consistently improves detection performance and robustness. Importantly, by treating intent as a conditioning signal rather than a shortcut, our approach mitigates reliance on superficial stylistic cues and ensures that verification remains evidence-centered.

In summary, we make three contributions:

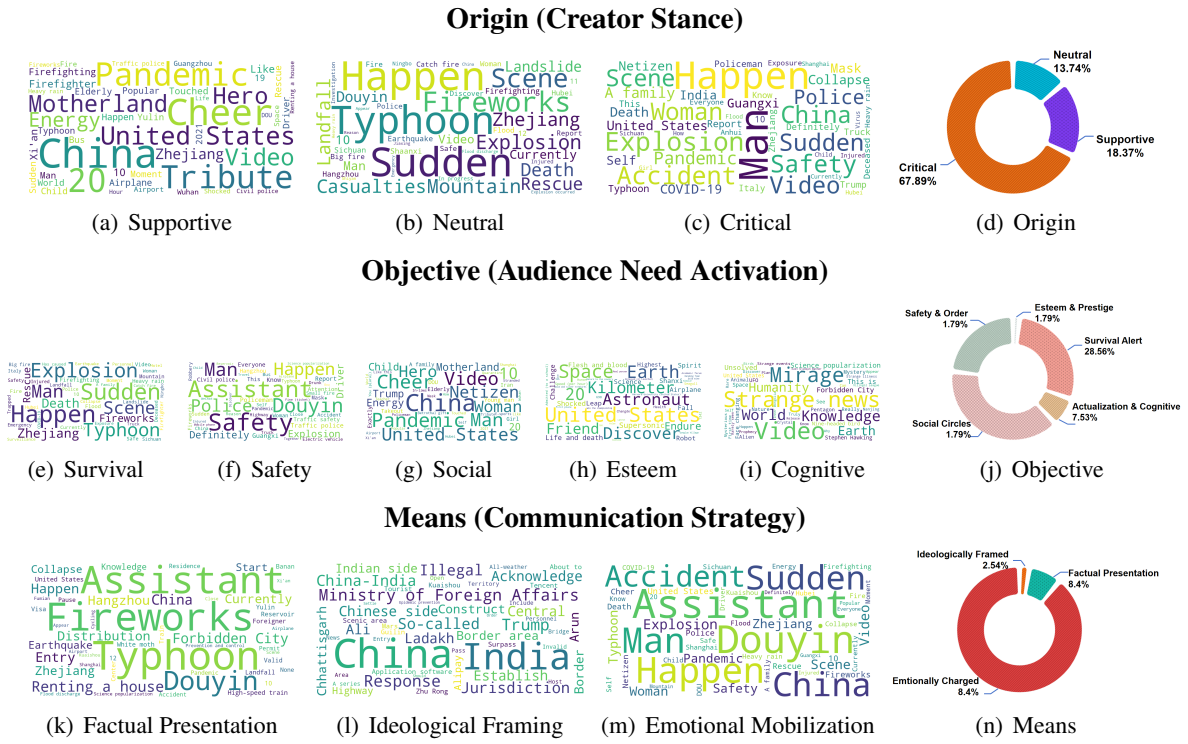
- **Framework**: We introduce OOM, a computationally explicit framework for modeling communicative intent in news videos.
- **Approach**: We show that treating intent as an intermediate semantic condition improves news video understanding via intent-conditioned reasoning and representation learning.
- **Empirical Insight**: We provide evidence that *audience need activation* plays a central role in news video framing, with implications for intent-aware video misinformation governance.

2 Related Work

2.1 News Intent Understanding

News intent has been studied as a communicative goal that shapes how information is selected and framed, rather than merely what is stated (Zhao et al., 2025; Maharana et al., 2022; Zeng et al., 2025a; Yang et al., 2024a,b; Pang et al., 2025; Yang et al., 2026; Lu and Li, 2026; Lu et al., 2025; Tong et al., 2024). Prior work often models intent as a latent cognitive state and attempts to decompose it using theories of intentional action. While conceptually informative, intent is typically represented in implicit or ad-hoc forms, limiting standardization and reuse. More recent frameworks (Wang et al., 2025; Wu et al., 2025) introduce structured intent elements within the news production process, but largely focus on text-based news and intent identification as an end task.

With the rise of visually rich news, intent analysis has been extended to image–text settings, highlighting the role of visual framing and affective cues. However, intent in these studies is still commonly treated as a task-specific label or latent variable, rather than a unified semantic abstraction that supports downstream reasoning.



(a) Supportive (b) Neutral (c) Critical (d) Origin

(e) Survival (f) Safety (g) Social (h) Esteem (i) Cognitive (j) Objective

(k) Factual Presentation (l) Ideological Framing (m) Emotional Mobilization (n) Means

Figure 2: Word clouds and label distributions for categories under the Origin–Objective–Means (OOM) framework. Word clouds (left) highlight distinct lexical patterns across Origin, Objective, and Means, while donut charts (right) show the overall class distributions for each intent dimension.

2.2 Multimodal and Video Intent Modeling

Beyond news, multimodal intent understanding has been explored in human-centered AI, including intent recognition, reasoning, and discovery from multimodal data (Zhang et al., 2025b). Related work addresses challenges such as modality heterogeneity and temporal misalignment, and frames intent understanding as an inference problem requiring contextual or commonsense reasoning(Li et al., 2023). Despite these advances, intent is generally treated as a prediction target, limiting its use for guiding reasoning or improving robustness.

2.3 Intent-Aware Misinformation Analysis

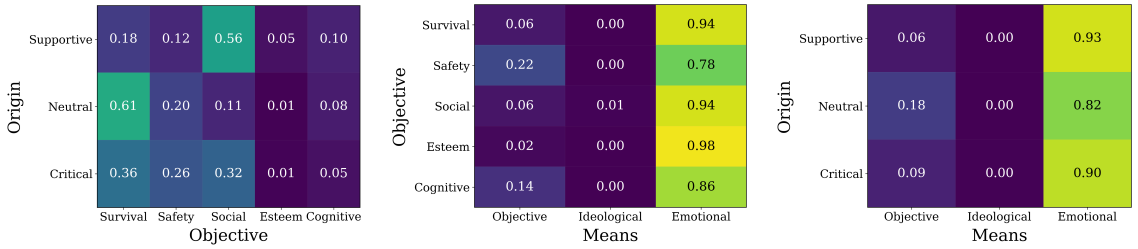
In misinformation analysis, intent-related cues are often linked to persuasion strategies and emotional framing. Recent studies incorporate such cues as auxiliary context and emphasize that intent should not directly determine veracity (Wu et al., 2025; Wang et al., 2025). Nevertheless, existing approaches rely on loosely defined or task-specific intent signals and lack a unified representation that separates communicative goals from factual correctness. In contrast, our work models intent as an explicit intermediate semantic representation that conditions multimodal reasoning while avoiding shortcut learning.

3 Dataset Construction and Analysis

3.1 Dataset Construction

We extend two widely used multimodal news video datasets, **FakeSV** and **FakeTT**, by performing large-scale annotation under the **Origin–Objective–Means (OOM)** framework. Each instance keeps its original veracity label and is additionally assigned an intent tuple $I = (I^{origin}, I^{objective}, I^{means})$, where Origin denotes the creator stance, Objective denotes the audience need being activated, and Means denotes the communication strategy.

Intent annotation process. To obtain reliable intent labels, each video was independently annotated by multiple trained annotators following detailed guidelines derived from the OOM definitions. Our annotation pipeline involved three primary annotators and two validation annotators. Annotators were instructed to judge the overall communicative framing of each video by considering all available cues rather than isolated signals. For each intent dimension, the final label was determined via majority voting. Cases with low inter-annotator agree-



(a) Origin → Objective (b) Objective → Means (c) Origin → Means

Figure 3: Row-normalized conditional co-occurrence heatmaps between OOM intent dimensions: (a) Origin→Objective, (b) Objective→Means, and (c) Origin→Means. Values indicate conditional probabilities $P(\cdot \mid \cdot)$, revealing non-uniform dependencies among intent components.

ment were further reviewed through discussion, and highly ambiguous samples were removed to ensure label quality. In total, tie cases accounted for 197 out of 5,616 samples (3.51%), and 53 samples (0.94%) were removed after discussion. The final validation agreement reached 100%. To further improve fairness and reduce systematic bias, annotators were selected from diverse research backgrounds. These results indicate strong annotation reliability.

3.2 Dataset Analysis

We analyze the annotated data from three complementary perspectives: (1) **lexical grounding** of intent categories, (2) **label distribution** across intent dimensions, and (3) **conditional co-occurrence** between intent components.

Lexical grounding and label distribution. Figure 2 summarizes both lexical and distributional characteristics of OOM labels. The word clouds (left) reveal distinct lexical patterns for categories under Origin, Objective, and Means, indicating that intent categories are well grounded in language and correspond to different framing styles and targeted needs. The donut charts (right) further show that intent labels are highly imbalanced across all three dimensions, reflecting dominant communicative patterns in real-world news videos (e.g., emotionally charged strategies and high-urgency needs appear more frequently).

Conditional co-occurrence (row-normalized). To quantify dependencies between intent components, we compute row-normalized conditional probabilities from contingency tables. Given two intent dimensions A (row variable) and B (column variable), let

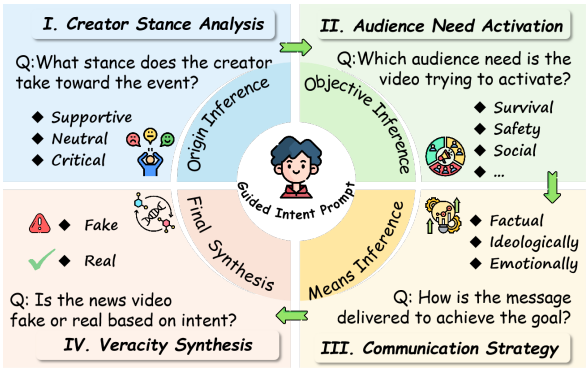


Figure 4: Overview of the Intent-Guided Prompting (IGP) framework for intent-conditioned reasoning.

$$C_{a,b} = \sum_{i=1}^{N} \mathbb{I}[A_i = a \wedge B_i = b] \tag{1}$$

denote the co-occurrence count (i.e., the contingency table entry) for category a in A and category b in B. We then compute the row-normalized conditional probability used in the heatmaps as

$$P(B = b \mid A = a) = \frac{C_{a,b}}{\sum_{b'} C_{a,b'}}, \tag{2}$$

where rows with $\sum_{b'} C_{a,b'} = 0$ are set to zero.[1]

In Figure 3, we instantiate Eq. (2) for the three pairs: (a) $A = \textit{Origin}, B = \textit{Objective}$, (b) $A = \textit{Objective}, B = \textit{Means}$, and (c) $A = \textit{Origin}, B = \textit{Means}$.

4 Method

4.1 Overview and Design Rationale

Our core design principle is to treat communicative intent as an explicit intermediate semantic repre-

[1]This exactly matches our implementation: we normalize each row of the crosstab and fill missing values with 0.0.

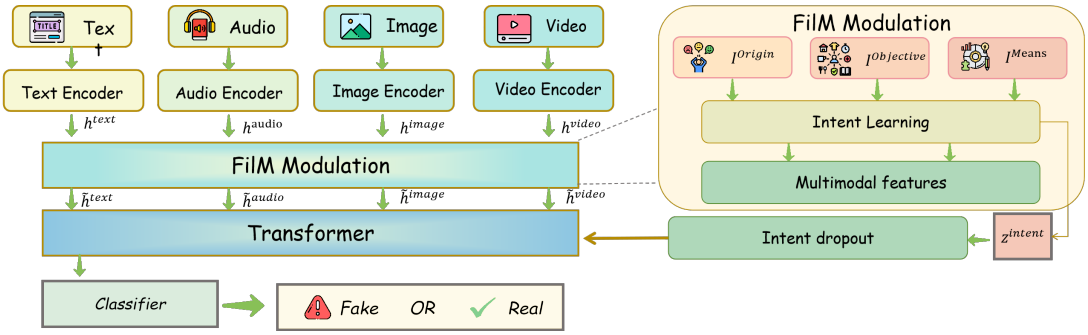


Figure 5: Intent-conditioned multimodal detection framework (ICMD).

sentation that conditions reasoning and evidence interpretation, rather than as a latent factor or a direct predictor of veracity. Under this principle, intent provides contextual information about why a news video is framed in a particular way, while factual correctness must still be determined by the underlying evidence. Based on this view, we develop two complementary methods that operationalize intent conditioning at different levels: (1) **Intent-Guided Prompting (IGP)** for intent-conditioned LLM reasoning, and (2) an **intent-conditioned multimodal detection network (ICMD)** for representation learning.

4.2 Method for News Intent Identification

Large language models (LLMs) enable scalable reasoning over complex multimodal content, but when communicative intent is treated as an implicit latent factor, LLMs may conflate stylistic cues with factual correctness, leading to shortcut-based or unstable predictions. To address this issue, we propose **Intent-Guided Prompting (IGP)**, a theory-guided prompting framework grounded in the Origin–Objective–Means (OOM) representation. As illustrated in Figure 4, IGP decomposes intent-conditioned reasoning into three stage-wise analyses (Origin, Objective, and Means) followed by a veracity synthesis step. Different from intent identification as a standalone task, IGP assumes that intent labels are given and uses them solely as structured semantic conditions for reasoning.

Problem Definition. Given a news video with associated textual content (e.g., title and description), let x denote the input and M an LLM. We assume that the communicative intent has been annotated or inferred in advance under the OOM framework:

$$\mathcal{I} = (I^{\text{origin}}, I^{\text{objective}}, I^{\text{means}}), \quad (3)$$

where I^{origin} denotes creator stance, $I^{\text{objective}}$ the audience need being activated, and I^{means} the communication strategy. The task is to predict veracity by reasoning over content *conditioned* on intent:

$$y = M(x, \mathcal{I} \mid p_{\text{guided}}), \quad y \in \{\text{fake}, \text{real}\}. \quad (4)$$

Origin-Guided Reasoning. The first stage conditions reasoning on the given *Origin* label. Rather than inferring stance, the model examines whether tone, evidence selection, and framing are consistent with the declared stance. Inconsistencies between strong evaluative language and weak evidence are treated as potential warning signals.

Objective-Guided Reasoning. The second stage conditions reasoning on the *Objective* label, i.e., the audience need being activated. The model evaluates whether needs such as survival alerts, safety concerns, social belonging, esteem, or cognitive curiosity are supported by verifiable evidence, helping identify exaggeration or emotional amplification.

Means-Guided Reasoning. The third stage conditions reasoning on the *Means* label. The model assesses whether the communication strategy (e.g., emotional mobilization or ideological framing) is appropriate given the available evidence.

Intent-Conditioned Veracity Synthesis. Finally, the model integrates the above analyses to produce a veracity judgment. Importantly, intent is not used as a decision rule but as a contextual lens that shapes how evidence is interpreted. The model is explicitly instructed to justify its decision based on evidence–intent alignment or mismatch.

4.3 Method for News Intent Application

We next introduce **ICMD**, our intent-conditioned multimodal detection framework, and how the identified intent is incorporated into a multimodal de-

Model	FakeSV				FakeTT			
	ACC	F1	M-P	M-R	ACC	F1	M-P	M-R
GLM-4.6V-9B	0.5541	0.1936	0.8248	0.1072	0.5891	0.4451	0.7338	0.2260
+IGP	0.6151	0.4896	0.7248	0.3696	0.6591	0.6451	0.8338	0.5260
LLaVA-OneVision-8B	0.5710	0.3014	0.8072	0.1853	0.5459	0.4504	0.9051	0.2998
+IGP	0.6126	0.5427	0.6614	0.4602	0.5895	0.5669	0.8211	0.4329
Qwen2.5-VL-32B	0.5820	0.3047	0.8997	0.1834	0.6732	0.6653	0.8381	0.5516
+IGP	0.6407	0.4825	0.8598	0.3354	0.7102	0.7048	0.8531	0.6274
Qwen2.5-VL-72B	0.6474	0.5569	0.7477	0.4436	0.7324	0.7563	0.8156	0.7050
+IGP	0.7087	0.7181	0.8097	0.6452	0.7691	0.7959	0.8298	0.7647
Qwen3-VL-32B	0.6865	0.6570	0.7244	0.6011	0.6792	0.7406	0.7070	0.7775
+IGP	0.6935	0.6903	0.6967	0.6840	0.6938	0.7496	0.7229	0.7783
InternVL3.5-38B	0.5844	0.3464	0.8077	0.2204	0.5013	0.2946	0.8889	0.1766
+IGP	0.6148	0.4339	0.8155	0.2956	0.5931	0.5405	0.8071	0.4063
InternVL3-78B	0.6084	0.4060	0.8377	0.2680	0.6777	0.6548	0.8865	0.5192
+IGP	0.6424	0.4941	0.8418	0.3497	0.6872	0.6719	0.8788	0.5439
GPT-5	0.6844	0.5464	0.7077	0.3204	0.7768	0.7937	0.8691	0.7302
+IGP	0.7094	0.6227	0.8825	0.4810	0.7842	0.7994	0.8821	0.7309

Table 1: Performance comparison on FakeSV and FakeTT. For each backbone model, the first row reports results under Direct prompting, while the indented row (IGP) reports results with intent-conditioned prompting.

tection model. Consistent with our design rationale, intent is treated as auxiliary semantic context rather than a predictive shortcut. As illustrated in Figure 5, ICMD injects the OOM intent into multimodal representation learning via FiLM-based feature modulation and intent-aware aggregation.

Problem Setting. Given a news video v with multimodal inputs $\mathcal{X} = \{X^{\text{text}}, X^{\text{audio}}, X^{\text{video}}, X^{\text{image}}\}$, and its OOM intent $\mathcal{I} = (I^{\text{origin}}, I^{\text{objective}}, I^{\text{means}})$, the goal of **ICMD** is to predict veracity while conditioning evidence interpretation on intent.

Structured Intent Representation. Each intent dimension is mapped to a learnable embedding:

$$\begin{aligned} e^{\text{origin}} &= f_{\text{origin}}(I^{\text{origin}}), \\ e^{\text{objective}} &= f_{\text{objective}}(I^{\text{objective}}), \\ e^{\text{means}} &= f_{\text{means}}(I^{\text{means}}), \end{aligned} \tag{5}$$

which are concatenated and projected to obtain a unified intent representation:

$$z^{\text{intent}} = f_{\text{intent}}([e^{\text{origin}}; e^{\text{objective}}; e^{\text{means}}]). \tag{6}$$

This representation captures communicative intent while remaining agnostic to factual correctness.

Intent-Conditioned Evidence Modulation. We extract modality-level evidence features $\{h^{\text{text}}, h^{\text{audio}}, h^{\text{video}}, h^{\text{image}}\}$. In ICMD, to inject intent without shortcut bias, we adopt feature-wise linear modulation (FiLM):

$$\tilde{h}^m = \gamma^m(z^{\text{intent}}) \odot h^m + \beta^m(z^{\text{intent}}), \tag{7}$$

allowing intent to dynamically reweight evidence across modalities.

Intent-Aware Aggregation and Training. The intent-conditioned features and intent token are treated as a token sequence and processed by a Transformer. The intent token aggregates cross-modal information and serves as the final representation for classification. To discourage shortcut reliance, we apply **intent dropout** during training by randomly masking the intent token with a fixed probability. We optimize ICMD using a standard **cross-entropy loss** for the final multi-class classification objective.

Overall, by modeling intent as a conditioning mechanism rather than a predictive feature, ICMD enables intent-aware yet evidence-centered multimodal fake news reasoning.

5 Experiments

5.1 Experimental Setup

Datasets. We conduct experiments on two multimodal fake news video benchmarks, **FakeSV** and

FakeTT. Each instance contains multimodal inputs and a binary veracity label. In addition, each video is associated with three intent labels under the Origin–Objective–Means (OOM) framework: creator stance, audience need activation, and communication strategy.

Evaluation Metrics. We report Accuracy (ACC) and F1-score (F1) as primary metrics. For completeness, we additionally report macro-precision (M-P) and macro-recall (M-R) where appropriate.

Implementation Details. For LLM-based experiments, we use fixed prompt templates with identical decoding settings across methods. For neural models, we follow standard training protocols with consistent optimization settings. All hyperparameters are kept the same across compared methods to ensure fair evaluation.

5.2 Experiment I: Intent-Guided Prompting (IGP)

This experiment evaluates whether explicit intent conditioning improves LLM-based multimodal fake news reasoning. Importantly, intent labels are provided as structured semantic conditions and are not used as predictors of veracity. For all prompting strategies, the model is provided with a fixed set of eight sampled video frames together with the textual content available in the dataset (e.g., title or description).

5.2.1 Backbone Models

We evaluate Intent-Guided Prompting (IGP) on a diverse set of vision–language backbones to test the generality of intent-conditioned prompting across model families and scales. Specifically, we include GLM-4.6V-9B (Cui et al., 2025), LLaVA-OneVision-8B, Qwen2.5-VL-32B, Qwen2.5-VL-72B (Bai et al., 2025), Qwen3-VL-32B (Yang et al., 2025), InternVL3.5-38B, InternVL3-78B (Zhu et al., 2025), and GPT5 (Hurst et al., 2024). For each backbone, we report results under Direct prompting (multimodal content only) and IGP, using identical decoding settings for fair comparison.

5.2.2 Main Results

Table 1 reports results across multiple vision–language backbones on FakeSV and FakeTT. Overall, IGP consistently improves upon Direct prompting, suggesting that introducing OOM intent as an explicit semantic condition provides both a principled and practical benefit for multimodal fake news reasoning. By decomposing intent into Origin-, Objective-, and Means-guided steps, IGP encourages evidence-centered verification and reduces reliance on superficial stylistic cues, leading to more stable and interpretable LLM judgments across backbones.

Prompt Strategy	FakeSV		FakeTT	
	ACC	F1	ACC	F1
Direct	0.6474	0.5569	0.7324	0.7563
CoT	0.6540	0.6380	0.7656	0.7881
VoT	0.6548	0.6658	0.7490	0.7920
IGP	0.7087	0.7181	0.7691	0.7959

Table 2: Comparison of different prompting strategies under Qwen2.5-VL-72B on FakeSV and FakeTT.

Intent Structure	FakeSV		FakeTT	
	ACC	F1	ACC	F1
w/o Origin	0.6742	0.6689	0.7134	0.7354
w/o Objective	0.6681	0.6530	0.6846	0.7014
w/o Means	0.6888	0.6758	0.7239	0.7333
IGP	0.7087	0.7181	0.7691	0.7959

Table 3: Ablation study on intent components under Qwen2.5-VL-72B.

5.2.3 Comparison of Prompting Strategies

Using the best-performing backbone (Qwen2.5-VL-72B), we further compare Direct, Chain-of-Thought (CoT), Verification of Thought (VoT), and Intent-Guided Prompting (IGP). Results in Table 2 show that while CoT and VoT provide moderate gains over Direct prompting, IGP yields the strongest and most stable improvements. This demonstrates that structured intent conditioning is complementary to generic reasoning prompts and provides additional benefits beyond reasoning depth alone.

5.2.4 Ablation on Intent Components

To assess the contribution of each intent dimension, we perform ablations by removing one component at a time. As shown in Table 3, removing any intent component consistently degrades performance, indicating that all three OOM dimensions provide complementary signals for intent-conditioned reasoning. Among them, removing *Objective* leads to the largest performance drop, suggesting that audience-need conditioning plays the most critical role in guiding evidence-centered LLM reasoning.

Table 4: Performance comparison of ICMD with traditional multimodal fusion methods and large vision-language models on the FakeSV and FakeTT datasets.

Method	FakeSV				FakeTT			
	ACC	M-F1	M-P	M-R	ACC	M-F1	M-P	M-R
TikTec	0.751	0.750	0.752	0.751	0.753	0.752	0.753	0.752
FANVN	0.750	0.750	0.751	0.750	0.762	0.746	0.762	0.742
Fact-R1	0.756	0.747	0.777	0.720	0.744	0.727	0.778	0.683
SV-FEND	0.793	0.792	0.796	0.793	0.802	0.790	0.804	0.785
FakingRec	0.796	0.796	0.797	0.796	0.793	0.786	0.782	0.781
MMVD	0.826	0.826	0.827	0.826	0.804	0.793	0.804	0.790
PNRN	0.833	0.833	0.836	0.833	0.817	0.817	0.818	0.811
ICMD	**0.838**	**0.837**	**0.837**	**0.838**	**0.829**	**0.823**	**0.828**	**0.821**

Variant	FakeSV		FakeTT	
	ACC	F1	ACC	F1
w/o Origin	0.827	0.827	0.814	0.807
w/o Objective	0.820	0.820	0.816	0.809
w/o Means	0.824	0.823	0.821	0.817
w/o FiLM	0.818	0.818	0.813	0.810
ICMD	0.838	0.837	0.829	0.823

Table 5: Ablation study of intent-conditioned multimodal detection framework.

5.3 Experiment II: Intent-Conditioned Multimodal Detection

We next evaluate whether communicative intent can improve multimodal fake news detection when incorporated as an explicit conditioning signal in representation learning, rather than as a predictive shortcut. All results are reported under five-fold cross-validation. For modality encoders, we use BERT (Radford et al., 2019) for text, VGG19 (Simonyan and Zisserman, 2014) for image (cover) features, C3D (Tran et al., 2015) for video features, and VGGish for audio features. We apply intent dropout with a rate of 0.2, which yields the best performance in our experiments.

5.3.1 Baselines

We compare our intent-conditioned model (ICMD) with representative multimodal fake news detection baselines spanning different modeling paradigms, including traditional multimodal fusion and detection methods (e.g., TikTec (Shang et al., 2021), FANVN (Choi and Ko, 2021), Fact-R1 (Zhang et al., 2025a), SV-FEND (Qi et al., 2023), FakingRec (Bu et al., 2024), MMVD (Zeng et al., 2024), and PNRN (Kong et al., 2025)). This comparison assesses whether intent conditioning provides benefits beyond backbone capacity or fusion design.

5.3.2 Main Results

Table 4 reports the main results on FakeSV and FakeTT. Overall, ICMD achieves the strongest or highly competitive performance across both datasets. Importantly, these gains are obtained by introducing intent as auxiliary semantic context to modulate evidence representations, rather than using intent as a direct indicator of veracity. The results demonstrate that explicit intent modeling can complement multimodal evidence and improve detection performance in a principled manner.

5.3.3 Ablation Study

To examine the role of intent in multimodal detection, we ablate individual intent components (w/o Origin, w/o Objective, w/o Means) as well as the intent-conditioned modulation mechanism (w/o FiLM). As shown in Table 5, removing any intent component or disabling FiLM-based conditioning degrades performance, confirming that both structured intent and intent-conditioned modulation are important. Among the three components, removing *Objective* causes the largest drop, underscoring the importance of audience-need conditioning.

5.4 Summary of Findings

Across both LLM-based reasoning and neural multimodal detection settings, we consistently observe that modeling communicative intent as an explicit intermediate semantic representation yields tangible benefits. Intent-conditioned prompting improves the stability and interpretability of LLM reasoning, while intent-conditioned representation learning enhances multimodal fake news detection without introducing shortcut bias. Together, these results provide empirical evidence that communicative intent serves as a valuable conditioning signal

for robust and principled multimodal fake news analysis.

6 Conclusion

We argue that robust multimodal news understanding requires modeling why content is framed, and we introduce Origin–Objective–Means (OOM) as an explicit intermediate intent representation. Large-scale annotation shows OOM labels are lexically grounded with imbalanced, structured dependencies. We further propose Intent-Guided Prompting (IGP) for intent-conditioned LLM reasoning and an intent-conditioned multimodal detection framework (ICMD) with feature-wise modulation to reduce shortcut reliance. Across two benchmarks and multiple backbones, intent conditioning consistently improves effectiveness and interpretability, supporting intent-aware intermediate semantics for multimodal misinformation analysis.

Limitations

Although our results demonstrate the value of modeling communicative intent as an intermediate semantic condition, this study still has several limitations. First, the proposed OOM framework relies on human annotation of Origin, Objective, and Means labels. Despite the use of detailed guidelines, multi-annotator labeling, majority voting, and additional review for low-agreement cases, intent interpretation remains inherently subjective and may vary across annotators, cultural backgrounds, and media contexts. Second, our experiments are conducted only on two short-form news video benchmarks, FakeSV and FakeTT. As a result, the generalizability of our findings to other domains, languages, platforms, and longer-form news videos remains to be verified. Third, both IGP and ICMD assume that intent labels are available or can be reliably inferred in advance. In real-world applications, errors in intent identification may propagate to downstream reasoning and veracity prediction. Finally, the OOM label distributions are notably imbalanced, and some of our experimental settings remain simplified, such as fixed frame sampling for LLM prompting and standard modality encoders for multimodal detection. These factors may limit robustness under distribution shift or more complex real-world conditions. Future work will explore more scalable automatic intent induction, broader cross-domain evaluation, and stronger temporal modeling for real-world deployment.

Ethics Statement

This paper adheres to the ACM Code of Ethics and Professional Conduct. Firstly, the dataset utilized does not contain sensitive private information and poses no harm to society. Secondly, proper attribution is given to relevant papers and the sources of pre-trained models, along with detailed references to the toolkits used. Furthermore, our code will be released under the license of any artifacts used. Lastly, the proposed fake news video detection method is designed to contribute to the safety and stability of the internet environment and public opinion. In addition, AI-assisted tools were used solely for grammar checking and language polishing, and did not affect the scientific content, experimental design, or conclusions of this work.

Acknowledgments

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101), the National Natural Science Foundation of China (No. 62272374, No. 62192781), the Natural Science Foundation of Shaanxi Province (No.2024JC-JCQN-62), the State Key Laboratory of Communication Content Cognition under Grant No. A202502, the Key Research and Development Project in Shaanxi Province (No. 2023GXLH-024), and the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

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

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. *arXiv preprint arXiv:2407.16670*.

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

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.

Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, et al. 2025. Glm-tts technical report. *arXiv preprint arXiv:2512.14291*.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Johan Edstedt, Amanda Berg, Michael Felsberg, Johan Karlsson, Francisca Benavente, Anette Novak, and Gustav Grund Pihlgren. 2022. Vidharm: A clip based dataset for harmful content detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1543–1549. IEEE.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.

Eungyeom Ha, Heemook Kim, and Dongbin Na. 2024. Hod: New harmful object detection benchmarks for robust surveillance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 183–192.

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*, pages 4684–4698.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Claire Wonjeong Jo, Magdalena Wojcieszak, et al. 2024. Harmful youtube video detection: A taxonomy of online harm and mllms as alternative annotators. *arXiv preprint arXiv:2411.05854*.

EunKyo Kang. 2019. Health information in the news media: Evaluating sources and subject of articles, and the intention to advertise. In *Proceedings of the 9th International Conference on Digital Public Health*, pages 121–121.

Xiangzheng Kong, Zhi Zeng, Chenxi Zhu, Zihan Ma, and Minnan Luo. 2025. Harmony in chaos: A progressive noise-resilient network for robust fake news video detection. *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11963–11974.

Wayne Lu and Yiheng Li. 2026. From blind transfer to wise selection: Prototype-driven neighbor-domain adaptation for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 818–826.

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

Zihan Ma, Minnan Luo, Yiran Hao, Zhi Zeng, Xiangzheng Kong, and Jiahao Wang. 2025. Bridging interests and truth: Towards mitigating fake news with personalized and truthful recommendations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 490–503.

Adyasha Maharana, Quan Hung Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal intent discovery from livestream videos. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 476–489.

Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological review*, 50(4):370.

Trisha Mittal, Sanjoy Chowdhury, Pooja Guhan, Snikitha Chelluri, and Dinesh Manocha. 2024. Towards determining perceived audience intent for multimodal social media posts using the theory of reasoned action. *Scientific Reports*, 14(1):10606.

Yucai Pang, Zhou Yang, Qian Li, Shihong Wei, and Yunpeng Xiao. 2025. Topic videolization: A rumor detection method inspired by video forgery detection technology. *IEEE Transactions on Knowledge and Data Engineering*, 37(6):3753–3765.

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.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Catriona Silvey. 2016. Speaking our minds: Why human communication is different, and how language evolved to make it special, by thom scott-phillips.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497.

Herun Wan, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, and Zhi Zeng. 2025. Difar: Enhancing multimodal misinformation detection with diverse, factual, and relevant rationales. *arXiv preprint arXiv:2508.10444*.

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Siyuan Ma, and Haonan Cheng. 2025. Exploring news intent and its application: A theory-driven approach. *Information Processing & Management*, 62(6):104229.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Computing Surveys*, 55(8):1–38.

Jiaying Wu, Fanxiao Li, Min-Yen Kan, and Bryan Hooi. 2025. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. *arXiv preprint arXiv:2505.15489*.

Liang Xiao, Qi Zhang, Chongyang Shi, Shoujin Wang, Usman Naseem, and Liang Hu. 2024. Msynfd: Multi-hop syntax aware fake news detection. In *Proceedings of the ACM web conference 2024*, pages 4128–4137.

Wenyan Xu, Dawei Xiang, Tianqi Ding, and Weihai Lu. 2025. Mmm-fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty. *arXiv preprint arXiv:2510.25120*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhou Yang, Yucai Pang, Qian Li, Shihong Wei, Rong Wang, and Yunpeng Xiao. 2024a. A model for early rumor detection base on topic-derived domain compensation and multi-user association. *Expert Systems with Applications*, 250:123951.

Zhou Yang, Yucai Pang, Xuehong Li, Qian Li, Shihong Wei, Rong Wang, and Yunpeng Xiao. 2024b. Topic audiolization: A model for rumor detection inspired by lie detection technology. *Information Processing & Management*, 61(1):103563.

Zhou Yang, Yucai Pang, Bin Yang, Haoyang Zhang, and Yunpeng Xiao. 2026. Dr-dgrnet: A model for intent-based disinformation recognition using dynamic graph representation learning. *IEEE Transactions on Computational Social Systems*.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in llms. *arXiv preprint arXiv:2405.08760*.

Zhi Zeng, Minnan Luo, Xiangzheng Kong, Hao Guo, Hui Yang, and Xiang Zhao. 2025a. Multimodal biases mitigation: Multimodal multi-granularity causal reasoning framework for multimodal fake news detection. In *2025 IEEE 10th International Conference on Data Science in Cyberspace (DSC)*, pages 327–334. IEEE.

Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. 2024. Mitigating world biases: A multimodal multi-view debiasing framework for fake news video detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6492–6500.

Zhi Zeng, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025b. Understand, refine and summarize: Multi-view knowledge progressive enhancement learning for fake news video detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9216–9225.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025c. Imol: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933.

Zhi Zeng, Yifei Yang, Jiaying Wu, Xulang Zhang, Xiangzheng Kong, Herun Wan, Zihan Ma, and Minnan Luo. 2026. From manipulation to mistrust: Explaining diverse micro-video misinformation for robust debunking in the wild. In *Proceedings of the ACM Web Conference 2026*, pages 7621–7632.

Fanrui Zhang, Dian Li, Qiang Zhang, Jun Chen, Gang Liu, Junxiong Lin, Jiahong Yan, Jiawei Liu, and Zheng-Jun Zha. 2025a. Fact-r1: Towards explainable video misinformation detection with deep reasoning. *arXiv preprint arXiv:2505.16836*.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. 2025b. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *IEEE Transactions on Multimedia*.

Liyuan Zhang, Yang Yajing, Yan Yang, Yong Liu, Zhongyan Gui, Ruofan Li, and Hao Fei. 2025c. Mfsvfnd: Multimodal fusion network for detecting fake news on short video platforms. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 2123–2127.

Jingwei Zhao, Yuhua Wen, Qifei Li, Minchi Hu, Yingying Zhou, Jingyao Xue, Junyang Wu, Yingming Gao, Zhengqi Wen, Jianhua Tao, et al. 2025. Deep learning approaches for multimodal intent recognition: A survey. *arXiv preprint arXiv:2507.22934*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.