MultiPress: A Multi-Agent Framework for Interpretable Multimodal News Classification

Tailong Luo[1], **Hao Li**[†,2], **Rong Fu**[3], **Xinyue Jiang**[4], **Huaxuan Ding**[4], **Yiduo Zhang**[4], **Zilin Zhao**[4], **Simon Fong**[3], **Guangyin Jin**[5], **Jianyuan Ni**[6]
[1]New York Institute of Technology [2]University of Arizona [3]University of Macau
[4]Peking University [5]Chang'an University [6]Juniata College

***Abstract*—With the growing prevalence of multimodal news content, effective news topic classification demands models capable of jointly understanding and reasoning over heterogeneous data such as text and images. Existing methods often process modalities independently or employ simplistic fusion strategies, limiting their ability to capture complex cross-modal interactions and leverage external knowledge. To overcome these limitations, we propose *MultiPress*, a novel three-stage multi-agent framework for multimodal news classification. *MultiPress* integrates specialized agents for multimodal perception, retrieval-augmented reasoning, and gated fusion scoring, followed by a reward-driven iterative optimization mechanism. We validate *MultiPress* on a newly constructed large-scale multimodal news dataset, demonstrating significant improvements over strong baselines and highlighting the effectiveness of modular multi-agent collaboration and retrieval-augmented reasoning in enhancing classification accuracy and interpretability.**

***Index Terms*—Multimodal News Classification, Multi-Agent Systems, Retrieval-Augmented Generation**

I. Introduction

The rapid expansion of multimedia news content on digital platforms has transformed how information is produced and consumed. Modern news articles frequently combine textual narratives with images, videos, and other modalities to deliver richer and more engaging stories. This multimodal nature introduces new challenges for automated news topic classification, a critical task for news aggregation, recommendation systems, and public opinion analysis.

Traditional text-based classification methods fall short in fully exploiting the complementary information provided by visual content. Meanwhile, straightforward multimodal fusion techniques often struggle with modality imbalance, noisy inputs, and limited integration of external knowledge. Recent advances in Multimodal Large Language Models (MLLMs) [1]–[8] have shown strong potential for understanding heterogeneous data. However, directly applying MLLMs to multimodal news classification remains challenging due to difficulties in extracting and fusing cross-modal features, knowledge latency, hallucination, and the need for complex reasoning that jointly integrates multimodal inputs with external knowledge sources [9]–[11].

To address these challenges, we propose *MultiPress*, a modular multi-agent framework for interpretable multimodal news classification. As illustrated in Figure 1, specialized agents first analyze textual and visual inputs to extract modality-specific features. A Retrieval-Augmented Generation (RAG) [12]–[14] agent then enriches the context by retrieving relevant external knowledge via vector search [15]–[18]. Subsequently, an online reasoning agent iteratively refines the understanding by dynamically invoking MLLMs and web search, enabling more accurate and context-aware classification. The resulting signals are integrated through a gated fusion mechanism that adaptively weighs multimodal features and retrieved knowledge, while a reward-driven evaluation agent provides feedback for iterative refinement.

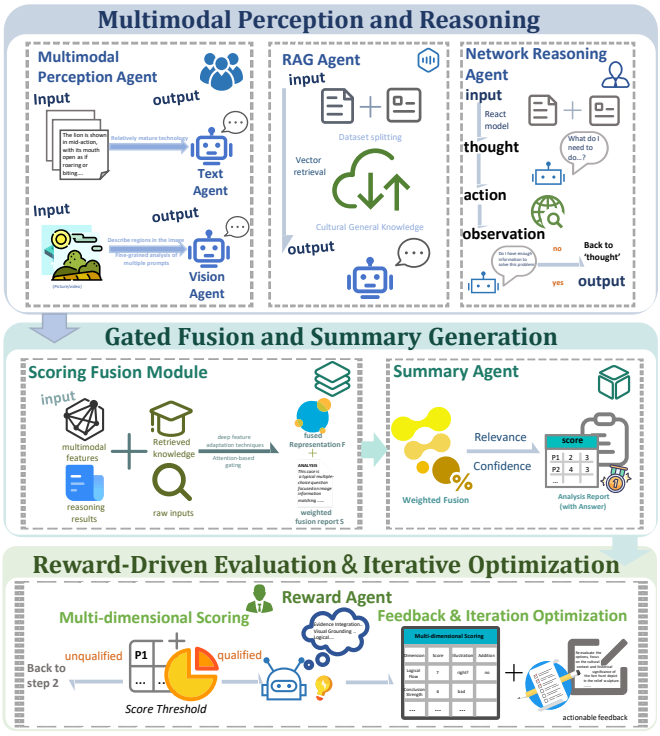

Fig. 1. Overview of our MultiPress Framework.

To rigorously evaluate our approach, we construct *NewsMM*, a large-scale multimodal news dataset comprising over 7,000 articles paired with images and annotated into eight high-level topic categories. This dataset reflects real-world multimodal news characteristics and serves as a robust benchmark.

Extensive experiments demonstrate that *MultiPress* significantly outperforms strong multimodal baselines. In addition to improving classification accuracy, *MultiPress* explicitly

† Corresponding author.

provides interpretable evidence—including textual cues, visual concepts, and retrieved knowledge—which is important for debugging, auditability, and user trust in real-world news applications.

We will make our dataset and code publicly available upon acceptance. Our main contributions are summarized as follows:

- We introduce *MultiPress*, a novel multi-agent framework integrating multimodal perception, retrieval-augmented reasoning, adaptive fusion, and reward-driven optimization for robust multimodal news classification.
- We construct *NewsMM*, a new large-scale multimodal news dataset with diverse topic categories.
- We conduct comprehensive experiments demonstrating that *MultiPress* significantly outperforms existing baselines, validating the effectiveness of modular multi-agent collaboration.

II. Method

This section introduces *MultiPress*, a three-stage multi-agent framework for interpretable multimodal news classification. Different from conventional multimodal classifiers that perform a single-pass fusion of text and images, MultiPress decomposes the task into collaborative agents specializing in perception, retrieval, reasoning, fusion, and reward-based refinement. Such modular decomposition enables stronger cross-modal grounding, reduces hallucination, and provides human-interpretable evidence for topic prediction. The overall workflow is illustrated in Figure 1.

A. Problem Formulation

Given a multimodal news instance consisting of textual content $\mathbf{T}$ (headline and body) and an associated image $\mathbf{I}$, the goal is to predict its topic label $y \in \{1, \ldots, C\}$, where $C = 8$ in *NewsMM*. In addition to classification accuracy, MultiPress explicitly produces an interpretable fusion report $\mathbf{S}$ grounded in multimodal cues and retrieved evidence:

$$(\hat{y}, \mathbf{S}) = \mathcal{M}(\mathbf{T}, \mathbf{I}), \tag{1}$$

where $\hat{y}$ is the predicted topic category and $\mathbf{S}$ contains supporting evidence, reasoning traces, and modality contributions.

B. Overview of the Multi-Agent Design

MultiPress consists of five specialized agents:

- **Perception Agent** extracts modality-specific semantic cues.
- **RAG Agent** retrieves external knowledge evidence.
- **Online Reasoning Agent** performs iterative ReAct-style reasoning with search.
- **Fusion Agent** integrates multimodal and retrieved evidence via gated scoring.
- **Reward Agent** evaluates report quality and triggers refinement.

Each agent communicates through structured messages rather than free-form generations. This design improves controllability and interpretability.

a) Agent Communication Protocol.: At each step, an agent receives a structured context:

$$\mathcal{C} = \{\mathbf{T}, \mathbf{I}, \mathbf{K}, \mathbf{R}, \mathbf{F}_{back}\}, \tag{2}$$

and outputs a structured message

$$m_a = \{\text{prediction}, \text{evidence}, \text{rationale}, \text{confidence}\}. \tag{3}$$

These messages are aggregated and passed across agents, enabling explicit modular collaboration.

C. Stage 1: Multimodal Perception and Knowledge-Augmented Reasoning

Stage 1 focuses on extracting modality-specific representations and enriching the context through retrieval and iterative reasoning.

1) Multimodal Perception Agent: The perception agent independently analyzes text and image inputs:

$$\mathbf{T}_{feat} = \mathcal{F}_T(\mathbf{T}), \quad \mathbf{I}_{feat} = \mathcal{F}_I(\mathbf{I}), \tag{4}$$

where $\mathcal{F}_T$ captures entities, events, sentiment polarity, and discourse cues, while $\mathcal{F}_I$ extracts objects, scenes, and visual attributes.

To enhance interpretability, the agent outputs structured cues:

$$\mathbf{Z}_T = \{\text{entities}, \text{keywords}, \text{text-summary}\}, \tag{5}$$

$$\mathbf{Z}_I = \{\text{objects}, \text{scene}, \text{image-summary}\}. \tag{6}$$

These intermediate representations provide explicit modality evidence for downstream fusion.

2) Retrieval-Augmented Generation (RAG) Agent: News classification often requires background knowledge that is not explicitly stated in the article, such as political figures, emerging events, or domain-specific terminology. Thus, the RAG agent retrieves evidence snippets from an external knowledge base:

$$\mathbf{K} = \mathcal{R}(\mathbf{T}, \mathbf{I}). \tag{7}$$

a) Cross-Modal Retrieval Scoring.: We embed both modalities into a shared latent space. Let $\mathbf{e}_T$ and $\mathbf{e}_I$ denote embeddings of $\mathbf{T}$ and $\mathbf{I}$, and $\mathbf{e}_j$ denote the embedding of a candidate knowledge item k_j. A retrieval score is computed as:

$$s_j = \beta \cos(\mathbf{e}_T, \mathbf{e}_j) + (1 - \beta) \cos(\mathbf{e}_I, \mathbf{e}_j), \tag{8}$$

where β balances textual and visual relevance. The top-k evidence items are returned:

$$\mathbf{K} = \{k_1, \ldots, k_k\}. \tag{9}$$

b) Knowledge Filtering.: To avoid noisy retrieval, we discard snippets with similarity below a threshold and prioritize recent evidence for time-sensitive news topics.

3) Online Reasoning Agent: After retrieval, the online reasoning agent performs iterative reasoning inspired by ReAct. At iteration t, the agent generates a thought, executes an action (search or verification), and updates observations:

$$\text{thought}_t = \phi(\mathbf{T}, \mathbf{I}, \mathbf{K}, \text{obs}_{t-1}), \tag{10}$$

$$\text{action}_t = \psi(\text{thought}_t), \tag{11}$$

$$\text{obs}_t = \omega(\text{action}_t). \tag{12}$$

The reasoning trace is summarized as:

$$\mathbf{R} = \{\text{step-wise rationale with citations}\}. \tag{13}$$

a) Stopping Criteria and Budget Control.: To ensure efficiency, reasoning terminates early if: (i) the predicted label distribution stabilizes across consecutive iterations, or (ii) the reward score exceeds threshold τ. We limit the loop to at most N iterations and M search actions per iteration.

D. Stage 2: Gated Fusion and Report Generation

Stage 2 integrates multimodal features and retrieved evidence into a unified representation.

1) Adaptive Gated Fusion: We compute modality-aware gating weights:

$$[\alpha_T, \alpha_I, \alpha_K, \alpha_R] = \text{softmax}(W_g[\mathbf{T}_{feat}; \mathbf{I}_{feat}; \mathbf{K}; \mathbf{R}]). \tag{14}$$

The fused representation is:

$$\mathbf{F} = \alpha_T \mathbf{T}_{feat} + \alpha_I \mathbf{I}_{feat} + \alpha_K \mathbf{K} + \alpha_R \mathbf{R}. \tag{15}$$

a) Reliability-Aware Fusion.: The gating mechanism implicitly reflects evidence reliability. For example, if the image is a generic stock photo or retrieval confidence is low, the model assigns smaller α_I or α_K, thereby mitigating modality noise.

2) Fusion Report Generation: A summary assistant generates the final interpretable report:

$$\mathbf{S} = \mathcal{H}(\mathbf{F}), \tag{16}$$

where $\mathbf{S}$ includes: predicted topic, supporting multimodal cues, retrieved evidence, and explanation.

E. Stage 3: Reward-Driven Evaluation and Iterative Refinement

Stage 3 introduces a reward agent that evaluates report quality and triggers refinement.

1) Reward Function: The reward is computed as:

$$r = \lambda_1 r_{cls} + \lambda_2 r_{ground} + \lambda_3 r_{cons}. \tag{17}$$

Here, r_{cls} measures classification confidence, r_{ground} evaluates evidence grounding, and r_{cons} measures cross-modal consistency. We set $\lambda_1 = 0.5, \lambda_2 = 0.3, \lambda_3 = 0.2$.

a) Reward Components.: r_{cls} is computed as the softmax margin between the top-1 and top-2 predicted topics. r_{ground} measures the proportion of rationale sentences that can be semantically matched to retrieved evidence. r_{cons} evaluates agreement between key entities extracted from text and objects extracted from images.

Algorithm 1 MultiPress Multi-Agent Inference Pipeline

Require: News text $\mathbf{T}$, image $\mathbf{I}$
Extract features $\mathbf{T}_{feat}, \mathbf{I}_{feat}$
Retrieve evidence $\mathbf{K} = \mathcal{R}(\mathbf{T}, \mathbf{I})$
for $t = 1$ to N **do**
  Online reasoning with search feedback $\mathbf{R}_t$
  Fuse evidence via gated fusion $\mathbf{F}_t$
  Generate report $\mathbf{S}_t = \mathcal{H}(\mathbf{F}_t)$
  Compute reward score $r_t = \mathcal{E}(\mathbf{S}_t)$
  if $r_t > \tau$ **then**
    break
  end if
end for
Output final prediction $\hat{y}$ and report $\mathbf{S}_t$

b) Iterative Refinement.: If $r < \tau$, the reward agent generates feedback:

$$\mathbf{F}_{back} = \mathcal{F}_{back}(r, \mathbf{S}), \tag{18}$$

which is sent back to the fusion stage for refinement until convergence.

F. Complexity Analysis

MultiPress introduces additional overhead due to retrieval and iterative reasoning. Let k be the retrieval depth and N the maximum reasoning iterations. The overall inference cost scales as:

$$\mathcal{O}(k + N \cdot M), \tag{19}$$

which remains practical for offline news classification and batch processing scenarios.

G. Summary

Algorithm 1 summarizes the full MultiPress inference procedure. Overall, MultiPress decomposes multimodal news classification into modular agent collaboration:

$$(\hat{y}, \mathbf{S}) = \mathcal{H}\Big(\mathcal{G}(\mathcal{F}_T, \mathcal{F}_I, \mathcal{R}, \mathcal{O})\Big), \tag{20}$$

enabling robust multimodal understanding, external knowledge grounding, and interpretable reasoning through reward-driven refinement.

III. The NewsMM Dataset

Data Collection. We construct *NewsMM* by crawling multimodal news articles from reputable online news portals and social media platforms. Each instance includes a news article's headline and body text paired with one or more relevant images. To ensure diversity, the dataset covers eight topic categories: *Politics, Economy, Technology, Sports, Entertainment, Health, Environment,* and *Science*.

Textual content is preprocessed to remove advertisements and unrelated metadata. Images are filtered to retain only those directly relevant to the news content. Articles lacking quality images are excluded to maintain dataset integrity.

Annotation Process. A team of 10 annotators with journalism and media expertise labels each sample with one of the eight topic categories based on combined textual and visual

content. Each sample is independently labeled by three annotators; majority voting determines the final label. Disagreements are resolved through expert review.

To enhance consistency and reduce bias, a GPT-based topic prediction model is used as a complementary annotator. Samples with conflicting human and GPT labels undergo further manual verification or are discarded if ambiguity persists.

Data Statistics. *NewsMM* contains 7,200 multimodal news samples evenly distributed across eight categories (900 per category). The dataset is split into 90% training (6,480 samples) and 10% testing (720 samples).

Textual data averages 45.2 tokens per headline and 320.7 tokens per article body. Images are standardized to 256×256 pixels, with an average of 1.8 images per article.

Knowledge Base Construction. To support retrieval-augmented reasoning, we build a comprehensive external knowledge base (KB) tailored for news classification, integrating:

- **Multimodal News Samples:** All *NewsMM* training samples with text, images, and labels.
- **External News Corpora:** Large-scale textual datasets such as AG News, Reuters-21578, and CNN.
- **Domain-Specific Knowledge:** Curated knowledge graphs and encyclopedic entries related to current events, organizations, and entities.

The KB is indexed using a vector database optimized for fast semantic retrieval. Textual and visual features are embedded into a shared latent space via pretrained multimodal encoders, enabling efficient cross-modal similarity search. This design allows the RAG agent to retrieve highly relevant knowledge snippets that enrich classification context.

a) Category definition.: We define eight high-level topics following common news taxonomies, where each label covers a broad yet coherent semantic scope. For example, *Politics* includes elections, diplomacy, and government policies; *Economy* covers markets, finance, trade, and macroeconomic indicators; *Technology* focuses on consumer tech, AI, and industrial innovation; and *Environment* includes climate events, conservation, and energy transitions. During annotation, we provide annotators with concise label guidelines and borderline examples to reduce ambiguity between closely related categories (e.g., *Science* vs. *Technology*, *Health* vs. *Science*).

b) Quality control and agreement.: To ensure annotation reliability, each sample is labeled by three annotators, followed by majority voting. We additionally conduct spot checks on randomly sampled subsets and perform guideline refinement when systematic confusion is observed. As an agreement signal, we compute inter-annotator consistency based on the proportion of samples with unanimous votes and those requiring adjudication. In practice, ambiguous instances often arise when images are generic (e.g., portraits or stock photos) or when the article narrative spans multiple domains; such cases are either resolved by expert review or removed if the topic remains uncertain.

TABLE I
PERFORMANCE COMPARISON ON *NewsMM* TEST SET.

Model	Acc.	MP	MR	MF1
GLM-4V-9B	70.5	69.8	69.2	69.5
mPLUG-7B	68.9	68.1	67.5	67.8
Qwen2.5-VL-3B (zero-shot)	62.3	61.0	60.5	60.7
Qwen2.5-VL-3B (fine-tuned)	66.8	66.0	65.5	65.7
Qwen2.5-VL-7B (zero-shot)	77.4	76.8	76.1	76.4
Qwen2.5-VL-7B (fine-tuned)	80.2	79.5	79.0	79.2
GPT-4o-20250513	79.6	79.0	78.5	78.7
MultiPress (GPT-4o)	**91.2**	**90.7**	**90.1**	**90.4**
MultiPress (Qwen2.5-VL-7B)	85.3	84.7	84.1	84.4

IV. EXPERIMENTS

A. Experimental Settings

We evaluate *MultiPress* on *NewsMM* using GPT-4o [19] and Qwen2.5-VL-7B-Instruct [20] as backbone multimodal models. We report overall accuracy (Acc.), macro-averaged precision (MP), recall (MR), and F1-score (MF1) to assess performance across all categories.

Unless otherwise specified, the multi-agent pipeline follows a unified configuration: (i) the perception agent extracts text/image cues, (ii) the retrieval agent queries a FAISS-based vector index with top-$k = 5$, (iii) the online reasoning agent performs up to 3 ReAct-style iterations with 2–3 search steps per iteration, and (iv) the reward agent evaluates outputs along accuracy, groundedness, and cross-modal consistency, with an acceptance threshold $\tau = 0.85$. All methods share the same train/test split for fair comparison.

B. Baseline Methods

We compare *MultiPress* with a broad set of strong multimodal baselines covering both open-source and proprietary models. Our primary baselines include GLM-4V-9B [21], mPLUG-7B [22], Qwen2.5-VL-3B/7B-Instruct (zero-shot and fine-tuned), and GPT-4o.

C. Overall Performance on NewsMM

Table I summarizes the main results on *NewsMM*. *MultiPress* consistently outperforms strong multimodal baselines. Notably, fine-tuning Qwen2.5-VL-7B improves zero-shot accuracy from 77.4% to 80.2%, but still trails *MultiPress (Qwen2.5-VL-7B)* by 5.1 points and *MultiPress (GPT-4o)* by 11.0 points. This indicates that the improvement is not merely from stronger backbones or fine-tuning, but from the *multi-agent decomposition* that explicitly enables knowledge grounding and iterative correction.

D. Generalization Beyond News Domain

To validate whether MultiPress generalizes beyond news classification, we also apply it on three multimodal reasoning benchmarks using Qwen2.5-VL-7B-Instruct as the backbone. MultiPress improves accuracy from 24.05 to 31.42 on MathVerse, from 54.00 to 61.35 on MMStar, and from 19.80 to 26.17 on MMMU. These consistent gains (roughly +6–7 points) suggest that the proposed agentic retrieval and reasoning mechanism provides general benefits for multimodal understanding, rather than overfitting to the news domain.

TABLE II
ABLATION STUDY ON *NewsMM* TEST SET.

Variant	Acc.	MF1
Full *MultiPress* (GPT-4o)	91.2	90.4
w/o Retrieval-Augmented Generation	86.5	85.7
w/o Online Reasoning Agent	87.1	86.3
w/o Gated Fusion	84.9	84.1
w/o Reward-Driven Optimization	88.3	87.5
Text-only input	79.4	78.6
Image-only input	72.8	71.9

TABLE III
INTERPRETABILITY EVALUATION ON 100 SAMPLED TEST INSTANCES (HIGHER IS BETTER).

Model	Faithfulness	Groundedness	Usefulness
Qwen2.5-VL-7B	3.42	3.15	3.38
GPT-4o-20250513	4.10	3.85	4.45
MultiPress (GPT-4o)	**4.65**	**4.72**	**4.58**

E. Ablation Study

Table II shows that each module contributes meaningfully to the final performance. Removing the RAG agent drops accuracy by 4.7 points, verifying that external knowledge grounding is crucial for long-tail entities and emerging events. Disabling the online reasoning agent causes a 4.1-point decrease, indicating that iterative search-and-reflection can correct initial misunderstandings or incomplete evidence. The largest degradation comes from removing gated fusion, showing that adaptive modality weighting is essential for suppressing noisy images and resolving modality conflicts. Finally, the reward-driven optimization contributes +2.9 points, suggesting that explicit multi-criteria evaluation helps refine both the prediction and the interpretability report.

F. Effect of Reward-Driven Iteration

We further examine how many refinement iterations are necessary. With 0 iterations (single pass), MultiPress achieves 87.5% Acc. and 86.8% MF1. Allowing 1 iteration improves performance to 90.1% Acc. / 89.3% MF1, and 2 iterations reaches the best result of 91.2% Acc. / 90.4% MF1. Using 3 iterations yields marginal change, indicating that most gains are obtained within the first 1–2 refinement rounds and additional iterations show diminishing returns. This observation supports our design choice of using a small iteration budget for practical deployment.

G. Retrieval Sensitivity Analysis

We study how retrieval depth influences performance. When retrieval is shallow (top-$k = 1$), MultiPress obtains 89.4% Acc. / 88.6% MF1. Increasing to top-$k = 3$ improves to 90.8% Acc. / 89.9% MF1, and top-$k = 5$ achieves the best result of 91.2% Acc. / 90.4% MF1. However, using top-$k = 10$ slightly degrades performance (90.5% Acc. / 89.7% MF1), suggesting that overly large retrieval depth may introduce irrelevant or conflicting evidence, increasing reasoning burden. Overall, MultiPress remains robust across a reasonable range of k, and benefits most from moderate retrieval depth.

TABLE IV
EFFICIENCY COMPARISON AVERAGED PER SAMPLE.

Method	#Calls	Tokens	Latency
GPT-4o (direct)	1.0	1,450	1.8s
MultiPress (GPT-4o)	6.2	5,820	8.5s

H. Interpretability Evaluation

Beyond accuracy, MultiPress explicitly targets interpretability. We randomly sample 100 test instances and ask three annotators to score model-generated rationales on a 1–5 Likert scale across faithfulness (to inputs), groundedness (to retrieved evidence), and usefulness (for human understanding). Table III shows that MultiPress yields substantially higher groundedness (4.72 vs. 3.85 for direct GPT-4o), confirming that retrieval and online reasoning reduce hallucination by encouraging evidence-backed explanations. Meanwhile, the improvement in faithfulness indicates that gated fusion helps align rationales with modality-specific cues rather than producing generic explanations.

I. Efficiency Analysis

MultiPress introduces additional overhead due to retrieval and iterative reasoning. As shown in Table IV, compared with direct prompting, MultiPress uses more model calls and tokens, increasing latency from 1.8s to 8.5s per sample. However, we note that the evaluation setting includes online search and multi-step refinement, which primarily targets improved robustness and interpretability rather than minimal latency. In practice, the framework is well-suited for offline/batch news processing, and future work can explore distillation or caching strategies to reduce inference cost.

V. RELATED WORK

A. Multimodal News Classification

Multimodal news classification benefits from integrating textual and visual cues, showing improvements over text-only methods [23]. While traditional multimodal methods use co-attention and joint embeddings, they struggle with incorporating external knowledge and performing complex reasoning.

B. Retrieval-Augmented Multimodal Models

Retrieval-Augmented Generation (RAG) enhances models by retrieving external knowledge, extending to multimodal settings for improved grounding and reducing hallucinations [12], [24]. Recent agentic RAG frameworks, such as HM-RAG and MultiRAG, better coordinate retrieval and reasoning across modalities [25], [26].

C. Multi-Agent Reasoning and Fusion

Multi-agent systems have been applied to multimodal reasoning by decomposing tasks into specialized agents. These systems enable effective knowledge synthesis and dynamic fusion, which enhances cross-modal reasoning [25]. Our work builds on this by integrating multimodal perception, retrieval, reasoning, fusion, and refinement into a unified framework for interpretable news classification.

VI. Conclusion

We present *MultiPress*, a novel three-stage multi-agent framework for multimodal news topic classification. By decomposing the task into specialized agents for multimodal perception, retrieval-augmented reasoning, gated fusion, and reward-driven optimization, *MultiPress* effectively captures complex cross-modal interactions and integrates external knowledge to improve accuracy and robustness. To support evaluation, we construct *NewsMM*, a large-scale multimodal news dataset with over 7,000 annotated samples across eight topics. Extensive experiments demonstrate *MultiPress* significantly outperforms strong baselines, validating the benefits of modular multi-agent collaboration and retrieval-augmented reasoning.

Future work will explore extending to additional modalities such as video and audio, incorporating real-time news streams, and investigating advanced reasoning strategies to further enhance multimodal news understanding.

References

[1] Zhen-guo Xi, Wen-hua Chen, Xin Guo, Wei He, Yi-hui Ding, Bei-da Hong, Cheng Zhang, Xu Wang, Hui Lin, Rui-nan Li, et al., "The Rise and Potential of Large Language Model Based Agents: A Survey," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023, pp. 14087–14110.

[2] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem, "Camel: Communicative agents for "mind" exploration of large language model society," 2023.

[3] Qingyun Wu, Gagan Bansal, Jie Zhang, Yiran Wu, Shaokun Li, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," arXiv preprint arXiv:2308.08155, 2023.

[4] Aman Madaan, Niket Tandon, Prakhar Gupta, Kevin Hall, Luyu Gao, Bodhisattwa Majumder, S. Nagpal, Chen-tsung Wu, Uri Alon, A. Srivatsan, et al., "Self-Refine: Iterative Refinement with Self-Feedback," in *Advances in Neural Information Processing Systems*, 2023, vol. 36.

[5] Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Ziwen Wang, Huaxuan Ding, Zhuo Cheng, Wenhao Cao, Zhiyuan Feng, et al., "Hssbench: Benchmarking humanities and social sciences ability for multimodal large language models," *arXiv preprint arXiv:2506.03922*, 2025.

[6] Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al., "Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes," *arXiv preprint arXiv:2510.19400*, 2025.

[7] Jiaxin Liu and Zhaolu Kang, "Reasonact: Progressive training for fine-grained video reasoning in small models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, vol. 40, pp. 7188–7196.

[8] Jian Gao, Richeng Xuan, Zhaolu Kang, Dingshi Liao, Wenxin Huang, Zongmou Huang, Yangdi Xu, Bowen Qin, Zheqi He, Xi Yang, et al., "Laobench: A large-scale multidimensional lao benchmark for large language models," *arXiv preprint arXiv:2511.11334*, 2025.

[9] Zhibin Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, H. T. Madabushi, and Pascale Fung, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 1877–1901.

[11] Zhaolu Kang, Junhao Gong, Qingxi Chen, Hao Zhang, Jiaxin Liu, Rong Fu, Zhiyuan Feng, Yuan Wang, Simon Fong, and Kaiyue Zhou, "Multimodal multi-agent empowered legal judgment prediction," *arXiv preprint arXiv:2601.12815*, 2026.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Myle Ott, Wen-tau Chen, Alexis Conneau, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 9459–9474.

[13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinlong Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Han, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023.

[14] Zhaolu Kang, Junhao Gong, Wenqing Hu, Shuo Yin, Kehan Jiang, Zhicheng Fang, Yingjie He, Chunlei Meng, Rong Fu, Dongyang Chen, et al., "Quanteval: A benchmark for financial quantitative tasks in large language models," *arXiv preprint arXiv:2601.08689*, 2026.

[15] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yu Yang, Lain Li, Weinan Zhang, Yexiang Wang, Liyuan He, Han-lin Li, and Jun Wang, "MetaGPT: Meta Programming for Multi-Agent Collaborative Framework," arXiv preprint arXiv:2308.00352, 2023.

[16] Andres M. Bran, Sam Cox, Oliver Schilter, and Philippe Schwaller, "ChemCrow: Augmenting large-language models with chemistry tools," arXiv preprint arXiv:2304.05376, 2023.

[17] Yingjie He, Zhaolu Kang, Kehan Jiang, Qianyuan Zhang, Jiachen Qian, Chunlei Meng, Yujie Feng, Yuan Wang, Jiabao Dou, Aming Wu, et al., "How order-sensitive are llms? orderprobe for deterministic structural reconstruction," *arXiv preprint arXiv:2601.08626*, 2026.

[18] Chenghua Zhu, Siyan Wu, Xiangkang Zeng, Zishan Xu, Zhaolu Kang, Yifu Guo, Yuquan Lu, Junduan Huang, and Guojing Zhou, "Edis: Diagnosing llm reasoning via entropy dynamics," *arXiv preprint arXiv:2602.01288*, 2026.

[19] OpenAI, "Gpt-4o system card," 2024.

[20] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin, "Qwen2.5-vl technical report," 2025.

[21] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," 2024.

[22] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou, "mplug-owl: Modularization empowers large language models with multimodality," 2024.

[23] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang, "N24News: A new dataset for multimodal news classification," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, Eds., Marseille, France, June 2022, pp. 6768–6775, European Language Resources Association.

[24] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen, "A survey of multimodal retrieval-augmented generation," 2025.

[25] Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma, "Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation," *arXiv preprint arXiv:2504.12330*, 2025.

[26] Mingyang Mao, Mariela M. Perez-Cabarcas, Utteja Kallakuri, Nicholas R. Waytowich, Xiaomin Lin, and Tinoosh Mohsenin, "Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding," 2025.