HANIA: Planner-Guided Multimodal Graph Evidence Selection for Grounded Question Answering

Zafar Ali[1], Asad Khan[1], Nimbeshaho Thierry[2], Nabila Amir[1], Adam A. Q. Mohammed[4] and Pavlos Kefalas[3,*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Portland Institute Nanjing, NJUPT, Nanjing, China
[3]Dashub & School of Informatics, Aristotle University of Thessaloniki, Greece
[4]School of Information Science and Engineering, Shandong University, Qingdao, China

Abstract

Multimodal question answering remains sensitive to noisy, incomplete, and weakly grounded evidence. Long unstructured contexts can introduce redundancy and encourage unsupported generation, while flat retrieval may overlook relations needed for multi-step reasoning. We present **HANIA**, a planner-guided multimodal graph framework for evidence-grounded question answering. HANIA processes the supplied image and text using a frozen vision–language model to extract concise question-relevant visual evidence with explicit abstention. It then constructs an input-grounded multimodal graph and applies a two-group finite-state planner to coordinate descriptive and relational evidence. Coverage-aware pruning retains a compact evidence set based on relevance, graph confidence, concept coverage, and modality diversity. The selected passages, visual statements, and graph triples are provided to a frozen instruction-tuned decoder. We evaluate HANIA on ScienceQA using answer accuracy, evidence-filtering quality, evidence-budget sensitivity, and efficiency. The results show that structured evidence planning and compact graph-guided retrieval can support competitive multimodal question answering without target-dataset fine-tuning or iterative retrieval. The code is available at https://github.com/Zafar-southeast/HANIA.

Keywords

Graph-enhanced LLMs, Multimodal KG-RAG, Planner-guided evidence control, Input-grounded graph retrieval, Multimodal question answering

1. Introduction

Multimodal question answering requires grounded reasoning over heterogeneous evidence, including text, images or diagrams, and relational knowledge. Knowledge-graph question answering (KG-QA) [1] and retrieval-augmented generation (RAG) [2] offer complementary strengths: graphs expose relations useful for multi-step reasoning, while retrieval helps identify relevant evidence. However, multimodal KG-assisted QA remains challenging under limited computational budgets because visual and textual information must be grounded, connected, and organized before generation. Extracted triples may be weakly aligned with visual content because of missing or misgrounded relations [3], while vision–language models can still produce unsupported answers even when relevant evidence is available [4].

Graph-aware RAG methods address part of this problem by organizing evidence through explicit relational structure. KG^{2}RAG performs graph-guided expansion and chunk organization [5], while KERAG combines subgraph retrieval with filtering and summarization [6]. Other approaches introduce explicit reasoning control, including PMR for progressive KGQA [7], SKRAG with finite-state reasoning skeletons [8], and GLOW with GNN–LLM integration for open-world KGQA [9]. Nevertheless, many existing methods remain text-centric, assume a pre-existing symbolic graph, or rely on complex training and retrieval pipelines. They provide limited support for constructing compact, question-aware graph evidence directly from supplied multimodal inputs.

Graph-enhanced LLMs for trustwOrthy Web data management (GLOW), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

*Corresponding author.
zafar_ali@seu.edu.cn (Z. Ali); asadkhan@seu.edu.cn (A. Khan); thierryn@pdx.edu (N. Thierry); nabilaameer1@gmail.com (N. Amir); adammohammed@sdu.edu.cn (A. A. Q. Mohammed); pavloskefalas@gmail.com (P. Kefalas)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

CEUR-WS.org/Vol-4313/glow-paper1.pdf

HANIA addresses this gap through input-grounded graph construction and planner-guided evidence control. Given a question and its supplied image and text, Qwen3-VL-Instruct [10] extracts concise question-relevant visual evidence with explicit abstention. Controlled rules and GLiREL [11] then extract question-aware relations from the visual and textual evidence. The resulting input-grounded graph preserves links between each retained statement or triple and its original source.

A two-group finite-state planner coordinates descriptive and relational evidence requirements. Coverage-aware pruning then retains a compact evidence set based on question relevance, graph confidence, concept coverage, and modality diversity. The selected passages, visual statements, and graph triples are serialized for a frozen Qwen3-8B-Instruct decoder [12]. HANIA therefore controls the evidence supplied to the decoder without target-dataset fine-tuning, decoder modification, external knowledge retrieval, or iterative retrieval. The main contributions are:

- We introduce **HANIA**, a planner-guided multimodal graph framework that constructs question-aware evidence directly from supplied images and text.
- We build an input-grounded multimodal graph with explicit provenance, preserving the connection between extracted relations and their supporting evidence.
- We propose a **two-group finite-state planner** with coverage-aware pruning and evaluate its effectiveness on ScienceQA using answer accuracy, evidence-filtering quality, evidence-budget sensitivity, and efficiency.

2. Related Work

Research on question answering has developed along four closely related directions: KG-based reasoning, multimodal QA, retrieval-augmented generation, and structured LLM reasoning. We briefly review these areas and position **HANIA** as a lightweight, planner-guided framework for constructing and selecting compact multimodal graph evidence.

KG-QA methods use structured triples to support interpretable reasoning [13]. Earlier systems relied on embedding-based retrieval and semantic parsing but often struggled with unseen entities, missing relations, and scalability. Recent approaches increasingly combine graph retrieval or structured reasoning with LLMs, including PMR [7], SKRAG [8], and GLOW [9]. These methods improve multi-hop reasoning through structured intermediate steps, but they are largely text-centric and commonly assume an existing symbolic graph. In contrast, **HANIA** constructs an instance-level multimodal graph directly from the supplied visual and textual evidence.

Multimodal QA requires coordinated reasoning over images, text, and relational information. Multimodal knowledge graph methods investigate how heterogeneous signals can support grounded reasoning [3]. MR-MKG applies graph neural message passing over multimodal nodes [14], while VaLiK studies automated vision-to-MMKG construction with alignment-noise mitigation [15]. VisDoM highlights the value of structured retrieval for visually rich inputs such as tables and diagrams [16], whereas VDocRAG preserves visual layout through image-based retrieval and modeling [17]. However, many systems still depend on coarse evidence aggregation or long multimodal contexts. **HANIA** instead converts the supplied context into compact question-aware graph evidence and organizes it through constrained planning. RAG improves factual consistency by grounding LLM outputs in retrieved evidence. Multimodal and graph-enhanced variants incorporate multimodal alignment [18, 2] and graph-based retrieval or organization, including KG^{2}RAG [5], KERAG [6], and Pythia-RAG [19]. More recent multimodal graph systems emphasize selective evidence control: GraphLoom uses reliability-calibrated graph evidence routing [20], while MKG-Attention applies budget-aware multimodal graph retrieval and evidence control for VideoQA [21]. Instruction tuning and Chain-of-Thought can improve reasoning control [22], while knowledge-augmented multimodal CoT provides structured intermediate reasoning [23]. **HANIA** differs by combining input-grounded multimodal graph construction, descriptive–relational planning, and coverage-aware pruning to produce a compact provenance-preserving evidence package for a frozen decoder, without external knowledge retrieval, decoder modification, or iterative retrieval.

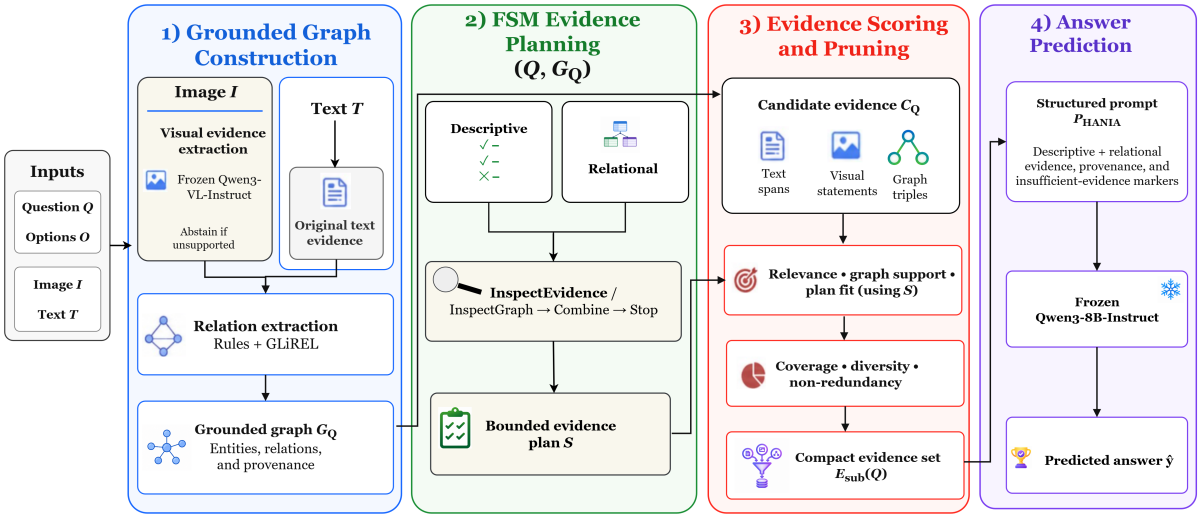


Figure 1: Architecture of HANIA. Input-grounded multimodal evidence is organized through finite-state planning and coverage-aware pruning before prediction by a frozen decoder.

3. Methodology: HANIA

Figure 1 presents **HANIA**, a planner-guided multimodal graph framework for evidence-grounded question answering. HANIA extracts question-relevant visual and textual evidence, constructs an input-grounded graph with explicit provenance, and uses a two-group finite-state planner to coordinate descriptive and relational requirements. Coverage-aware pruning then selects a compact set of passages, visual statements, and graph triples for a frozen QWEN3-8B-INSTRUCT decoder. All pretrained components are frozen, with no external knowledge retrieval.

3.1. Input-Grounded Evidence and Graph Construction

Let Q denote the question, $\mathcal{O}$ the answer options, and $X = (I, T)$ the supplied image and textual context, where either modality may be absent. For an available image, a frozen QWEN3-VL-INSTRUCT model extracts a concise, question-relevant statement:

$$D_Q^I = f_{\mathrm{VL}}(I, Q, \mathcal{O})\,. \tag{1}$$

The model reports only visibly supported information and returns an abstention statement when the requested evidence cannot be determined. The original text T is retained without generative rewriting. HANIA then constructs a question-aware graph $G_Q = (V_Q, E_Q, \pi_Q)$, where V_Q contains entities, properties, and numerical values from D_Q^I and T, E_Q contains the retained relations, and π_Q links each graph element to its supporting visual statement or text span.

A question-aware relation vocabulary $\mathcal{R}_Q$ is selected based on the required evidence type, including attributes, counts, comparisons, and spatial relations. Deterministic templates convert regular visual statements into triples, while GLIREL [11] extracts relations from textual evidence:

$$(\hat{r}_{jk}, c_{jk}) = f_{\mathrm{rel}}(e_j, e_k, s; \mathcal{R}_Q)\,, \tag{2}$$

where s is the supporting evidence unit and c_{jk} is the relation confidence. A relation is retained only if its entities are valid, $\hat{r}_{jk} \in \mathcal{R}_Q$, $c_{jk} \geq \tau_r$, and it is relevant to Q. Deterministic relations receive unit confidence. No external commonsense triples are added; unsupported requirements remain unresolved for the planning stage.

3.2. FSM-Constrained Two-Group Planning

Given Q and the input-grounded graph G_Q, HANIA builds a bounded evidence plan before pruning. The required evidence is organized into two groups:

$$\mathcal{P} = \{\textsc{Descriptive}, \textsc{Relational}\}. \tag{3}$$

The descriptive group covers attributes, colors, materials, counts, and shapes, while the relational group covers spatial, comparative, and object-level relations.

The plan is an ordered sequence $\mathfrak{S} = (s_1, \ldots, s_L)$, where $s_h = (a_h, g_h, c_h)$ and $L \leq H$. Here, a_h is the planner action, $g_h \in \mathcal{P}$ is the active evidence group, and c_h contains relevant entity, relation, or property cues. The action space is:

$$\mathcal{A}_{\text{plan}} = \{\textsc{InspectEvidence}, \textsc{InspectGraph}, \textsc{Combine}, \textsc{Stop}\}. \tag{4}$$

At each step, a frozen Qwen3-8B-Instruct planner [12] proposes the next state under finite-state constraints:

$$s_h = \text{Plan}_{\text{FSM}}(Q, G_Q, s_{<h}), \qquad (a_{h-1}, a_h) \in \mathcal{T}_{\text{FSM}}, \tag{5}$$

where $\mathcal{T}_{\text{FSM}}$ defines valid transitions. The controller rejects invalid transitions, avoids repeated inspection of the same group, and stops after at most H steps. Simple questions may use one evidence group, while more complex questions may combine both. The planner identifies evidence requirements but does not generate the final answer.

3.3. Planner-Guided Retrieval and Coverage-Aware Pruning

The candidate pool $\mathcal{C}_Q$ contains textual spans, visual statements, and graph triples, each linked to its source through π_Q. Each candidate $e \in \mathcal{C}_Q$ is scored as:

$$R(e) = \alpha_1 S_{\text{q}}(e, Q) + \alpha_2 S_{\text{graph}}(e) + \alpha_3 S_{\text{plan}}(e, \mathfrak{S}), \qquad \sum_{i=1}^{3} \alpha_i = 1, \tag{6}$$

where S_{q} measures question relevance, S_{graph} uses retained relation confidence or direct source support, and S_{plan} rewards alignment with an active planner group. Scores are normalized to $[0, 1]$, and no scorer is fine-tuned on ScienceQA.

A fixed top-K ranking may retain redundant evidence while missing required concepts. HANIA therefore applies greedy coverage-aware pruning. Given the selected set $\mathcal{A}$, the marginal value of a candidate is:

$$\Delta(e \mid \mathcal{A}) = R(e) + \beta_1 \Delta_{\text{cov}}(e \mid \mathcal{A}) + \beta_2 \Delta_{\text{div}}(e \mid \mathcal{A}) - \beta_3 \text{Redundancy}(e, \mathcal{A}), \tag{7}$$

where Δ_{cov} rewards newly covered question concepts and planner groups, while Δ_{div} rewards complementary modalities or evidence types. Starting from $\mathcal{A}_0 = \varnothing$, HANIA selects:

$$e_t^* = \arg\max_{e \in \mathcal{C}_Q \setminus \mathcal{A}_{t-1}} \Delta(e \mid \mathcal{A}_{t-1}), \qquad \mathcal{A}_t = \mathcal{A}_{t-1} \cup \{e_t^*\}. \tag{8}$$

Selection stops when the budget $N_{\max}$ is reached or the best remaining score falls below τ. The final package is $\mathcal{E}_{\text{sub}}(Q) = \mathcal{A}_T$, where T is the stopping step. If a required planner group remains unsupported, HANIA inserts an insufficient-evidence marker instead of retaining weak evidence.

3.4. Structured Evidence Prompting and Answer Generation

HANIA serializes the retained evidence $\mathcal{E}_{\text{sub}}(Q)$ while preserving the provenance links defined by π_Q. Evidence is grouped according to the descriptive and relational requirements identified by the planner:

$$\mathcal{F}^{(g)} = \text{Serialize}\left(\{(e, \pi_Q(e)) \mid e \in \mathcal{E}_{\text{sub}}(Q),\ \text{group}(e) = g\}\right), \quad g \in \{\text{desc}, \text{rel}\}. \tag{9}$$

Each item contains its original text or visual statement, any associated graph triple, and its supporting modality. If reliable support is missing, an insufficient-evidence marker is inserted. For answer options $\mathcal{O} = \{o^{(j)}\}_{j=1}^{n}$, the structured decoder prompt is:

$$P_{\text{HANIA}} = \left[Q; \mathcal{O}; \mathcal{F}^{(\text{desc})}; \mathcal{F}^{(\text{rel})}\right]. \tag{10}$$

A frozen Qwen3-8B-Instruct decoder [12] selects an answer using only the supplied evidence. Let $\{y_i^{(j)}\}_{i=1}^{L_j}$ denote the tokens of option $o^{(j)}$. Each option is scored using length-normalized log-likelihood:

$$S\left(o^{(j)}\right) = \frac{1}{L_j} \sum_{i=1}^{L_j} \log p_\theta \left(y_i^{(j)} \mid y_{<i}^{(j)}, P_{\text{HANIA}}\right), \qquad \hat{y} = \arg\max_j S\left(o^{(j)}\right). \tag{11}$$

This constrained formulation avoids unrestricted generation and ensures that the prediction matches one of the provided options.

3.5. Frozen-Backbone Configuration

HANIA keeps the vision–language, relation-extraction, planner, and answer models frozen; none is fine-tuned on ScienceQA. Its contribution therefore lies in question-aware evidence construction, provenance-preserving graph representation, finite-state planning, and coverage-aware pruning. The relation-confidence threshold, planner horizon, evidence budget, prompts, and scoring weights are selected on the validation split and fixed before testing. Gold answers are used only for validation and evaluation and remain unavailable during graph construction, planning, pruning, and prediction.

4. Experimental Evaluation

We evaluate **HANIA** on ScienceQA using answer accuracy, evidence-filtering quality, evidence-budget sensitivity, and end-to-end latency. All hyperparameters are selected on the validation split and fixed before test evaluation.

4.1. Dataset and Evaluation Protocol

ScienceQA [24] is a multimodal multiple-choice benchmark whose instances may contain an image, textual context, both, or neither. We use the official test split and only the supplied context, without external knowledge retrieval. Answer accuracy is the primary metric; evidence precision, recall, and F1 use candidate-level binary reference relevance labels fixed across all re-run methods and used only for evaluation. We also vary the evidence budget k and report end-to-end latency. All Table 1 baselines are our re-run implementations on identical test instances and matched inference conditions. With deterministic decoding, each system is run once; 95% confidence intervals use 10,000 paired bootstrap resamples of test instances. McNemar's test with Holm correction is used for the pre-specified comparisons with Qwen3-VL Direct, Flat-RAG, and SelF-Reasoner.

4.2. Overall Results and Component Ablation

Table 1 reports the controlled comparison and component-level ablations on ScienceQA. All methods use the same evaluation protocol, and evidence-filtering metrics are computed against the same relevance annotations.

Comparative analysis. HANIA achieves the highest accuracy of 87.80%, exceeding Qwen3-VL Direct by 2.40 points (95% CI: $[1.95, 2.85]$, Holm-adjusted $p < 0.003$). Compared with Flat-RAG, the closest matched baseline, HANIA improves accuracy by 0.90 points (95% CI: $[0.52, 1.28]$, Holm-adjusted $p = 0.008$) and F1 by 2.86 points (95% CI: $[2.10, 3.50]$). It also exceeds SelF-Reasoner by 0.56 accuracy points (95% CI: $[0.18, 0.94]$, Holm-adjusted $p = 0.021$).

Table 1
Controlled comparison and component-level ablation results on ScienceQA. All values are percentages.

Model/Variant	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑
Baseline comparison				
Qwen3-VL Direct [10]	85.40	–	–	–
FOCUS + LLaVA-1.5-13B [25]	74.40	57.20	51.80	54.37
EfficientLLaVA [26]	83.05	61.90	57.40	59.56
Cantor (Gemini) [27]	84.96	63.10	58.80	60.87
SciTune-ScienceQA (CTOM, 7B) [28]	86.11	64.30	59.70	61.91
SelF-Reasoner (UnifiedQA-Large) [29]	87.24	66.20	60.50	63.22
Flat-RAG	86.90	65.30	59.80	62.43
Component ablation				
w/o Graph	86.70	66.50	61.00	63.65
w/o Coverage-Aware Pruning	86.40	65.20	62.00	63.58
w/o Insufficient-Evidence Markers	86.30	66.00	61.80	63.84
HANIA	**87.80**	**68.57**	**62.31**	**65.29**

Table 2
Effect of the retained-evidence budget k on ScienceQA. All metrics except latency are percentages.

Budget	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	Latency (s/q) ↓
$k = 1$	82.60	**75.40**	43.80	55.40	**1.72**
$k = 3$	86.10	71.20	57.40	63.56	2.18
$\mathbf{k = 5}$	**87.80**	68.57	62.31	65.29	2.64
$k = 7$	86.70	64.10	67.20	**65.61**	3.09
$k = 9$	85.30	59.80	**70.40**	64.68	3.51

Ablation analysis. Removing each component reduces performance. Replacing coverage-aware pruning with fixed top-k ranking causes the largest precision drop, indicating that marginal-gain selection limits redundant evidence. Removing insufficient-evidence markers produces the largest accuracy reduction, while removing the graph lowers both accuracy and evidence F1. Together with the Flat-RAG comparison, these results show that HANIA benefits from provenance-linked graph construction, constrained planning, coverage-aware pruning, and explicit handling of unsupported evidence.

4.3. Filtering Quality and Evidence-Budget Selection

We vary the retained-evidence budget k while keeping all other settings fixed. Precision and recall are computed against the reference relevance labels for candidate evidence items, and F1 is their harmonic mean. Latency denotes the mean end-to-end processing time per question, including visual evidence extraction, graph construction, planning, pruning, and answer prediction.

Increasing k improves recall but gradually reduces precision. Accuracy peaks at 87.80% with $k = 5$, while $k = 7$ produces the highest filtering F1 but slightly lower answer accuracy. At $k = 9$, the additional recall is accompanied by lower precision and accuracy, suggesting that redundant or weakly relevant evidence can distract the decoder. Mean latency also increases from 1.72 to 3.51 seconds per question. The value $k = 5$ was selected using the validation split before test evaluation because it provided the best validation accuracy with a moderate evidence and latency budget. The test-set analysis is therefore reported as a sensitivity study rather than a basis for parameter selection.

4.4. Controlled Analysis

Qwen3-VL Direct compares the complete HANIA pipeline with direct multimodal prompting under matched test instances and inference conditions. Flat-RAG uses the same candidate pool and frozen decoder but omits two-group planning and uses fixed top-k relevance ranking. The ablations respectively remove graph triples/graph support, replace greedy coverage-aware pruning with fixed top-k, or omit the insufficient-evidence marker. As reported in Section 4.2, removing any of these components reduces answer accuracy or evidence-filtering quality. The evidence-budget analysis further examines the trade-off among accuracy, filtering quality, and latency. Together, these results indicate that HANIA's gains arise from the combined contribution of provenance-linked graph evidence, constrained planning, coverage-aware selection, and explicit handling of unsupported evidence.

4.5. Parameter Selection

All parameters were selected on the ScienceQA validation split and fixed before test evaluation. The retained-evidence budget was set to $k = N_{\max} = 5$, the relation-confidence threshold to $\tau_r = 0.50$, the evidence stopping threshold to $\tau = 0.20$, and the planner horizon to $H = 4$. The relevance-scoring weights in Equation (6) were set to $\boldsymbol{\alpha} = (0.50, 0.25, 0.25)$ for question relevance, graph support, and planner alignment, respectively. The pruning weights in Equation (7) were set to $\boldsymbol{\beta} = (0.30, 0.20, 0.25)$ for concept coverage, evidence diversity, and redundancy. Planner decoding used temperature 0, and all prompt templates, thresholds, pretrained components, and evidence-processing rules remained unchanged across test experiments. The test split was not used for prompt design, hyperparameter tuning, or model selection.

4.6. Implementation Details

HANIA is implemented in PyTorch using Hugging Face Transformers. All pretrained components remain frozen. Each ScienceQA instance is processed using only its supplied question, answer options, image, and textual context. The frozen Qwen3-VL-Instruct model extracts a concise question-relevant visual statement, while the original textual context is retained without generative rewriting. Rule-based span extraction identifies entity candidates, and deterministic templates together with GLiREL construct provenance-linked relations. Low-confidence, malformed, and question-irrelevant relations are removed. The descriptive–relational planner uses frozen Qwen3-8B-Instruct and is applied once before coverage-aware pruning and prediction by the frozen Qwen3-8B-Instruct decoder. Planner decoding uses temperature 0 and is restricted to valid finite-state actions. The answer decoder assigns length-normalized scores only to the provided answer options. Experiments are conducted on a single NVIDIA A100 40 GB GPU. HANIA uses no external knowledge retrieval, iterative retrieval, decoder modification, or target-dataset fine-tuning.

5. Conclusion and Future Work

We presented **HANIA**, a planner-guided multimodal graph framework for evidence-grounded question answering. HANIA constructs question-aware graph evidence from the supplied visual and textual context, coordinates descriptive and relational requirements through a finite-state planner, and applies coverage-aware pruning before prediction with a frozen decoder. On ScienceQA, HANIA achieved 87.80% accuracy, 68.57% evidence precision, 62.31% recall, and 65.29% F1. The evidence-budget analysis identified $k = 5$ as the best setting for answer accuracy with moderate latency. Controlled comparisons and component-level ablations showed that provenance-linked graph construction, constrained planning, coverage-aware pruning, and explicit handling of unsupported evidence jointly improve answer accuracy and evidence quality over direct multimodal prompting and flat relevance ranking, without target-dataset fine-tuning or decoder modification. The study is limited to one benchmark and does not yet include detailed graph-quality, faithfulness, or question-type analyses. Future work will

evaluate HANIA on additional multimodal benchmarks, improve visual relation extraction and planner constraints, and conduct deeper faithfulness, error, and graph-quality analyses.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) for grammar and spelling checking, paraphrasing and rewording, and formatting assistance. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Jiang, K. Zhou, X. Zhao, J.-R. Wen, UniKGQA: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph, in: Proceedings of the 11th International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.

[2] W. Chen, H. Hu, X. Chen, P. Verga, W. Cohen, MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, UAE, 2022, pp. 5558–5570.

[3] S. Zheng, W. Wang, J. Qu, H. Yin, W. Chen, L. Zhao, MMMKGR: Multi-hop multi-modal knowledge graph reasoning, in: Proceedings of the 39th IEEE International Conference on Data Engineering (ICDE), Anaheim, CA, 2023, pp. 96–109.

[4] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, L. Bing, Mitigating object hallucinations in large vision-language models through visual contrastive decoding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13872–13882.

[5] X. Zhu, Y. Xie, Y. Liu, Y. Li, W. Hu, Knowledge graph-guided retrieval augmented generation, in: Proceedings of NAACL 2025, 2025, pp. 8912–8924.

[6] Y. Sun, K. Sun, Y. E. Xu, X. Yang, X. L. Dong, N. Tang, L. Chen, Kerag: Knowledge-enhanced retrieval-augmented generation for advanced question answering, in: Findings of EMNLP 2025, 2025, pp. 6194–6216.

[7] Y. Ren, Z. Nie, R. Zhang, X. Liu, Progressive multi-hop reasoning for question answering over knowledge graphs, Information Systems 140 (2026) 102721.

[8] X. Xu, Y. Wang, Y. Liu, S. Li, Skrag: A retrieval-augmented generation framework guided by reasoning skeletons over knowledge graphs, in: Findings of EMNLP 2025, 2025, pp. 13983–13994.

[9] H. Abdallah, I. Abdelaziz, P. Kalnis, E. Mansour, Leveraging LLM-GNN integration for open-world question answering over knowledge graphs, in: V. Demberg, K. Inui, L. Marquez (Eds.), Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 575–592. doi:10.18653/v1/2026.eacl-long.26.

[10] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al., Qwen3-vl technical report, arXiv preprint arXiv:2511.21631 (2025).

[11] J. Boylan, C. Hokamp, D. G. Ghalandari, GLiREL - generalist model for zero-shot relation extraction, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 8230–8245.

[12] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

[13] L. Liu, Y. Chen, M. Das, H. Yang, H. Tong, Knowledge graph question answering with ambiguous query, in: Proceedings of the 2023 ACM Web Conference (WWW), Austin, TX, 2023, pp. 2477–2486.

[14] J. Lee, Y. Wang, J. Li, M. Zhang, Multimodal reasoning with multimodal knowledge graph, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Bangkok, Thailand, 2024, pp. 10767–10782.

[15] J. Liu, S. Meng, Y. Gao, S. Mao, P. Cai, G. Yan, Y. Chen, Z. Bian, D. Wang, B. Shi, Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning, in: Proceedings of ICCV, 2025, pp. 981–992.

[16] M. Suri, P. Mathur, F. Dernoncourt, K. Goswami, R. A. Rossi, D. Manocha, VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation, in: Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, Albuquerque, NM, 2025, pp. 6088–6109.

[17] R. Tanaka, T. Iki, T. Hasegawa, K. Nishida, K. Saito, J. Suzuki, Vdocrag: Retrieval-augmented generation over visually-rich documents, in: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 24827–24837.

[18] B. Chen, A. Khare, G. Kumar, A. Akula, P. Narayana, Seeing beyond: Enhancing visual question answering with multi-modal retrieval, in: Proceedings of the 31st International Conference on Computational Linguistics (COLING), Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 410–421.

[19] Z. Ali, Y. Huang, A. Khan, G. Qi, Y. Zhang, J. Feng, C. Deng, P. Kefalas, Pythia-rag: Retrieval-augmented generation over a unified multimodal knowledge graph for enhanced qa, Knowledge-Based Systems 335 (2026) 115200. doi:10.1016/j.knosys.2025.115200.

[20] Z. Ali, A. Khan, A. Malik, P. Kefalas, Graphloom: Reliability-calibrated graph evidence routing for multimodal kg-rag, 2026. arXiv:2608.15056.

[21] A. Khan, Z. Ali, G. Qi, K. P. Bhuyan, P. Kefalas, Mkg-attention: Budget-aware multimodal graph retrieval for faithful video question answering, Neurocomputing 704 (2026) 134854. doi:10.1016/j.neucom.2026.134854.

[22] H. Liu, Z. Teng, L. Cui, C. Zhang, Q. Zhou, Y. Zhang, Logicot: Logical chain-of-thought instruction tuning, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 2908–2921.

[23] D. Mondal, S. Modi, S. Panda, R. Singh, G. S. Rao, KAM-CoT: Knowledge augmented multimodal chain-of-thoughts reasoning, in: Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI) and 36th Conference on Innovative Applications of Artificial Intelligence (IAAI) and 14th Symposium on Educational Advances in Artificial Intelligence (EAAI), Vancouver, Canada, 2024, pp. 18798–18806.

[24] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to explain: multimodal reasoning via thought chains for science question answering, in: Proceedings of the 36th International Conference in Neural Information Processing Systems (NeurIPS), New Orleans, LA, 2022.

[25] S. Jiang, C. Zhou, Y. Zhang, Y. Jin, Z. Liu, Fast or slow? integrating fast intuition and deliberate thinking for enhancing visual question answering, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 525–534.

[26] Y. Liang, Z. Wang, X. Xu, J. Zhou, J. Lu, EfficientLLaVA: Generalizable auto-pruning for large vision-language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 9445–9454.

[27] T. Gao, P. Chen, M. Zhang, C. Fu, Y. Shen, Y. Zhang, S. Zhang, X. Zheng, X. Sun, L. Cao, R. Ji, Cantor: Inspiring multimodal chain-of-thought of mllm, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 9096–9105.

[28] S. Horawalavithana, S. Munikoti, I. Stewart, H. Kvinge, K. Pazdernik, SCITUNE: Aligning large language models with human-curated scientific multimodal instructions, in: Proceedings of the 1st Workshop on NLP for Science (NLP4Science), Association for Computational Linguistics, Miami, FL, USA, 2024, pp. 58–72.

[29] Y. Wu, Z. Zhang, H. Zhao, Mitigating misleading chain-of-thought reasoning with selective filtering, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11325–11340.