Hybrid-DMKG: A Hybrid Reasoning Framework over Dynamic Multimodal Knowledge Graphs for Multimodal Multihop QA with Knowledge Editing

Li Yuan[1,2], Qingfei Huang[1,2], Bingshan Zhu[3], Yi Cai[1,2*], Qingbao Huang[4], Changmeng Zheng [5], Zikun Deng[1,2], Tao Wang[6]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2] Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[3] School of Big Data and Artificial Intelligence, Guangdong University of Finance & Economics
[4] School of Electrical Engineering, Guangxi University, Nanning, Guangxi, China
[5] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[6] Department of Biostatistics & Health Informatics, King's College London, London, United Kingdom
{seyuanli@mail,ycai@,zkdeng@}.scut.edu.cn, qbhuang@gxu.edu.cn, changmeng.zheng@polyu.edu.hk, tao.wang@kcl.ac.uk

Abstract

Multimodal Knowledge Editing (MKE) extends traditional knowledge editing to settings involving both textual and visual modalities. However, existing MKE benchmarks primarily assess final answer correctness, neglecting the quality of intermediate reasoning and robustness to visually rephrased inputs. To address this limitation, we introduce MMQAKE, the first benchmark for multimodal multihop question answering with knowledge editing. MMQAKE evaluates: (1) a model's ability to reason over 2–5-hop factual chains that span both text and images, including performance at each intermediate step; (2) robustness to visually rephrased inputs in multihop questions. Our evaluation shows that current MKE methods often struggle to consistently update and reason over multimodal reasoning chains following knowledge edits. To overcome these challenges, we propose Hybrid-DMKG, a hybrid reasoning framework built on a dynamic multimodal knowledge graph (DMKG) to enable accurate multihop reasoning over updated multimodal knowledge. Hybrid-DMKG first uses a large language model to decompose multimodal multihop questions into sequential sub-questions, then applies a multimodal retrieval model to locate updated facts by jointly encoding each sub-question with candidate entities and their associated images. For answer inference, a hybrid reasoning module operates over the DMKG via two parallel paths: (1) relation-linking prediction; (2) RAG Reasoning with large vision-language models. A background-reflective decision module then aggregates evidence from both paths to select the most credible answer. Experimental results on MMQAKE show that Hybrid-DMKG significantly outperforms existing MKE approaches, achieving higher accuracy and improved robustness to knowledge updates.

Introduction

With the rapid advancement and widespread adoption of large language models (LLMs) (Zhao et al. 2023; Achiam et al. 2023; Chang et al. 2024; Shen et al. 2025b,a; Zhang et al. 2025b,a; Wang et al. 2025b), knowledge editing (KE) has emerged as a critical research area (Meng et al. 2022a,b; Mitchell et al. 2022). KE aims to revise inaccurate, incomplete, or outdated knowledge encoded in LLMs while minimizing unintended alterations to unrelated content. To systematically evaluate whether such edits improve model responses, particularly for complex queries whose answers depend on the updated knowledge, Zhong et al. (2023) proposed the *multihop question answering with knowledge editing* (MQUAKE) task. MQUAKE requires models to perform multihop reasoning over modified knowledge and has been explored in text-only settings (Gu et al. 2024a; Shi et al. 2024; Lu et al. 2025). However, many real-world applications involve multimodal information, such as text, images, and videos, which present new challenges in fusing and representing diverse modalities (Zhang et al. 2024; Huang et al. 2024). This highlights the necessity of multimodal knowledge editing (MKE), an extension of KE that enables reasoning and modification both visual and textual modalities.

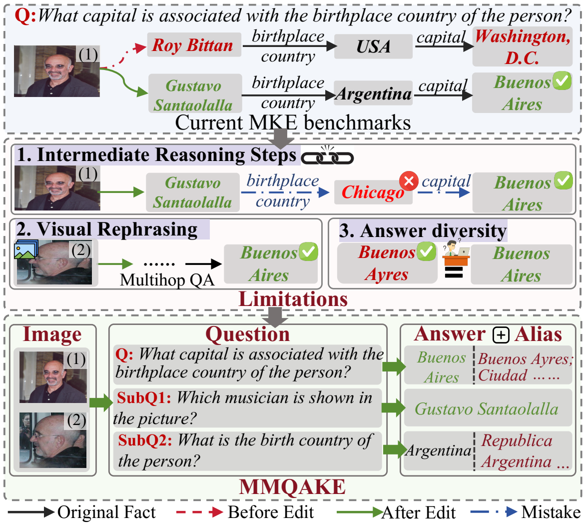


Figure 1: An example of our benchmark (MMQAKE), which differs in evaluation from existing MKE benchmarks.

Despite recent progress in MKE (Huang et al. 2024; Du

---

[*] Corresponding author: Yi Cai (ycai@scut.edu.cn)

Benchmark	Multimodal	Visual Rephrasing	Evaluation		
			Multihop Accuracy	Hop-wise Accuracy	Aliases
MQUAKE	✗	✗	✓	✓	✓
VLKEB(multihop)	✓	✗	✓	✗	✗
MMQAKE	✓	✓	✓	✓	✓

Table 1: The comparison of different benchmarks across various evaluation dimensions for multimodal multihop QA. "Visual Rephrasing" refers to the use of alternative images of the same entity to evaluate multihop reasoning. "Aliases" refers to whether answer aliases are accepted as correct during evaluation.

et al. 2025), current benchmarks primarily evaluate the correctness of final answers produced by large vision-language models (LVLMs) (Li et al. 2023; Liu et al. 2023; Zhu et al. 2024; Wang et al. 2025a; Chen, Wang, and Zhang 2025; Liang et al. 2025a; Cui et al. 2025), while giving little attention to the quality of intermediate reasoning and robustness to visually rephrased inputs. For example, in Figure 1, although the person's name is modified from "*Roy Bittan*" to "*Gustavo Santaolalla*", existing benchmarks still assess only the final answer "*Buenos Aires*" for the multihop question Q, without examining the reasoning steps required to derive it. Such end-only evaluation risks masking reasoning errors (Zhong et al. 2023), thereby limiting the reliability and interpretability of MKE performance. These issues manifest in three key limitations, as illustrated in Figure 1. (1) **Lack of accurate evaluation of intermediate reasoning steps.** In multihop question answering, models may occasionally produce the correct final answer while relying on outdated or incorrect facts (Gu et al. 2024a), such as (*Gustavo Santaolalla*, "**birthplace country**", *Chicago*). Ignoring the correctness of intermediate reasoning steps obscures the model's actual reasoning process and undermines the reliability of the evaluation. (2) **Lack of robustness evaluation under visual rephrasing.** Robust MKE methods should produce consistent outputs even when input images are visually modified (e.g., from image (1) to (2)). However, existing benchmarks often overlook this aspect, limiting the model's ability to generalize to real-world scenarios where visual content may be modified or presented in diverse forms. (3) **Neglect of valid alias diversity.** For example, answers such as *Buenos Ayres* are not recognized as equivalent to *Buenos Aires*, despite their semantic equivalence. This can penalize correct answers, undermining fair evaluation and potentially misrepresenting model performance.

To address these limitations, we propose Multimodal Multihop Question Answering with Knowledge Editing (MMQAKE), an extension of the VLKEB benchmark (Huang et al. 2024), as shown in Figure 1. MMQAKE features multihop questions requiring 2 to 5 reasoning steps, each aligned with a factual link in a reasoning chain. When multimodal knowledge is updated, models need to correctly propagate the revised information and generate answers that reflect the updated facts (Zhong et al. 2023). Besides, we evaluate predictions at each intermediate step (Gu et al. 2024a), enabling fine-grained assessment of reasoning quality. Additionally, we include visually rephrased images to test robustness to visual variations. Finally, following the MQUAKE evaluation protocol, we consider all valid aliases of the ground-truth answer (e.g., *Buenos Aires* and *Buenos Ayres*), as retrieved from Wikidata. The key differences between MMQAKE and existing benchmarks, including VLKEB and MQUAKE, are summarized in Table 1. Using MMQAKE, we further evaluate several representative MKE approaches to assess their effectiveness in complex reasoning scenarios. Our results reveal that many existing methods (Chen et al. 2020; Zhu et al. 2020; De Cao, Aziz, and Titov 2021; Mitchell et al. 2022; Zheng et al. 2023), struggle with the multihop and cross-modal challenges.

To address the faithfulness of current MKE methods in multihop question answering, we propose **Hybrid-DMKG**: a hybrid reasoning framework built upon dynamic multimodal knowledge graphs (DMKG). The DMKG represents knowledge as structured triples $(head, relation, tail)$, where entities are linked with corresponding images, and supports dynamic updates to accommodate evolving knowledge. This framework enriches semantic connections and enhances reasoning capabilities in LVLMs (Liang et al. 2025b; Li, Miao, and Li 2024). Moreover, inspired by Chain-of-Thought reasoning (Wei et al. 2022) and multihop question decomposition (Zhong et al. 2023; Gu et al. 2024a), we employ LLMs without fine-tuning to decompose multihop question into a sequence of sub-questions. For visual-based sub-questions, we utilize a multimodal retrieval model that jointly encodes the sub-question, candidate entities, and their associated images from the DMKG, with the goal of retrieving the entity most relevant to the sub-question as the answer. For reasoning-based sub-questions, we propose a hybrid reasoning module that operates along two parallel pathways to generate candidate answers: (1) relation-link prediction, which traverses the DMKG to infer an answer directly, and (2) retrieval-augmented generation–enhanced reasoning in the LVLM, which incorporates context from the DMKG. A background-reflective decision module then aggregates evidence from both paths to select the most credible answer. Our main contributions can be summarized as follows:

- We propose **MMQAKE**, the first benchmark for multimodal multihop question answering with knowledge editing, extending the existing MKE tasks. MMQAKE challenges models to reason over both textual and visual modalities across 2 to 5-hop factual chains. In addition, it evaluates robustness to visual rephrasing in multihop questions, simulating real-world scenarios where knowledge must be accurately updated and reflected through complex reasoning.

Datasets	Edit Number	2-hop	3-hop	4-hop	5-hop	Sub-question Number	Average Aliases
(MMQAKE) (Eval)	1,278	1,278	1,238	1,193	1,110	11,773	9.49

Table 2: Statistics of the MMQAKE dataset. The "Average Aliases" denotes the average number of answer aliases.

- We propose **Hybrid-DMKG**, a step-by-step reasoning framework built on a dynamic multimodal knowledge graph that continuously maintains and updates structured multimodal knowledge. By integrating complementary reasoning strategies, symbolic relation traversal, and retrieval-augmented generation in LVLM, this framework enhances the accuracy of multihop inference. Moreover, we propose a reflective decision module that effectively reconciles differing reasoning outputs, leading to more robust and faithful answers.
- Extensive experiments on MMQAKE with multimodal knowledge editing methods reveal that most struggle with multihop and cross-modal reasoning. Our proposed Hybrid-DMKG approach significantly outperforms existing baselines, demonstrating higher accuracy and improved robustness to knowledge updates.

Methodology

Problem Definition

Multimodal knowledge editing is formalized as a quadruple $\mathcal{D} = (x, v, o, \tilde{o})$, where x is the textual input, v is corresponding a image, and the objective is to update a fact from o to $\tilde{o}$. This editing operation is denoted as $f = (x, v, o \rightarrow \tilde{o})$. Based on this formulation, we introduce the task of **MMQAKE**, referred to as the textual MQUAKE task.

Given a multihop question Q associated with an image v, answering Q requires executing a sequence of intermediate queries that form a multihop reasoning chain. This process can be represented as: $C = [\{o, r_1, y_1\}, \{t_2, r_2, y_2\}, \dots, \{t_n, r_n, y_n\}]$. At the k-th hop, t_k denotes the subject, r_k the relation, and y_k the object. Notably, the object of the $(k-1)$-th fact serves as the subject of the k-th fact, i.e., $y_{k-1} = t_k$. In Figure 1, the initial set of relationships includes: ([IMAGE], *name*, *Roy Bittan*), (*Roy Bittan* , *birthplace country*, *USA*), and (*USA* , *capital*, *Washington, D.C.*). Based on this chain, a 3-hop question such as *What capital is associated with the birthplace country of the person?* can be formulated. When multimodal facts in the chain are edited ([IMAGE], *name*, *Roy Bittan* → *Gustavo Santaolalla*), LVLM leverages the updated knowledge to answer the multihop question correctly $y_n \rightarrow \tilde{y}_n$ (*Washington, D.C.* → *Buenos Aires*).

Besides, we argue that an effective MKE method should incorporate all edits from the knowledge corpus C into the model (Gu et al. 2024a), thereby enabling internal reasoning over the updated information. To evaluate whether these edits have been integrated, we assess the model's ability to answer decomposed sub-questions derived from multihop queries, as illustrated in Figure 1. The model needs to correctly answer each sub-question to ensure consistency throughout the reasoning chain, i.e., $((y_1, y_2, \dots, y_{n-1}) \rightarrow (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_{n-1}))$. To further evaluate generalization in the visual modality, we test the model on both the original image v and a visually rephrased image $\tilde{v}$ to evaluate the model's robustness and generalization across related visual inputs under the edited knowledge setting.

Dataset Construction

MMQAKE extends VLKEB (Huang et al. 2024) by evaluating models on **each step of multihop questions**, **visual rephrasing**, and **linguistic diversity in answers**, thereby increasing the complexity of reasoning depth and cross-modal understanding. Specifically, each multihop question in MMQAKE is augmented with three *paraphrased questions* generated via the ChatGPT API to simulate natural language ambiguity (Gu et al. 2024a; Lu et al. 2025). Additionally, each question is *decomposed into sub-questions*, each accompanied by paraphrases and annotated intermediate answers, enabling a *step-by-step evaluation* of the model's reasoning process. To evaluate robustness to visual rephrasing, we introduce alternative images from the VLKEB that depict the same entity as the original edited image. Finally, to ensure fair and semantically robust evaluation, we construct *answer alias sets* based on Wikidata references, mitigating the impact of linguistic variation in answers. Dataset statistics are summarized in Table 2.

Hybrid-DMKG Framework

Hybrid-DMKG is a parameter-preserving framework, which comprises the following key components: **(a) Dynamic MKG Construction:** We construct and maintain a structured DMKG, where knowledge is encoded as image-related triples. This structure enables efficient updates and deletions, supporting real-time adaptation to evolving facts. **(b) Question Decomposition:** We utilize an LLM to decompose multihop questions into multiple single-hop sub-questions, distinguishing between visual sub-questions that require image-based support and reasoning sub-questions that rely on structured knowledge. **(c) Cross-Modal Entity Retrieval from DMKG:** Visual sub-questions are handled using a cross-modal retrieval model that jointly encodes the visual query and each entity in the DMKG. The model then retrieves the most relevant entity as the answer. **(d) DMKG-Guided Hybrid Reasoning:** For reasoning sub-questions, candidate answers are generated through two parallel pathways: (1) relation linking within the DMKG to identify relevant answers, (2) an RAG-based method that enhances the LVLM's generation using DMKG-derived context. Then, a reflective decision module jointly evaluates the supporting background knowledge retrieved from the DMKG for each candidate and selects the most plausible answer.

Dynamic MKG Construction We use an MKG $\mathcal{G}$ as an external source to manage multimodal knowledge, provid-

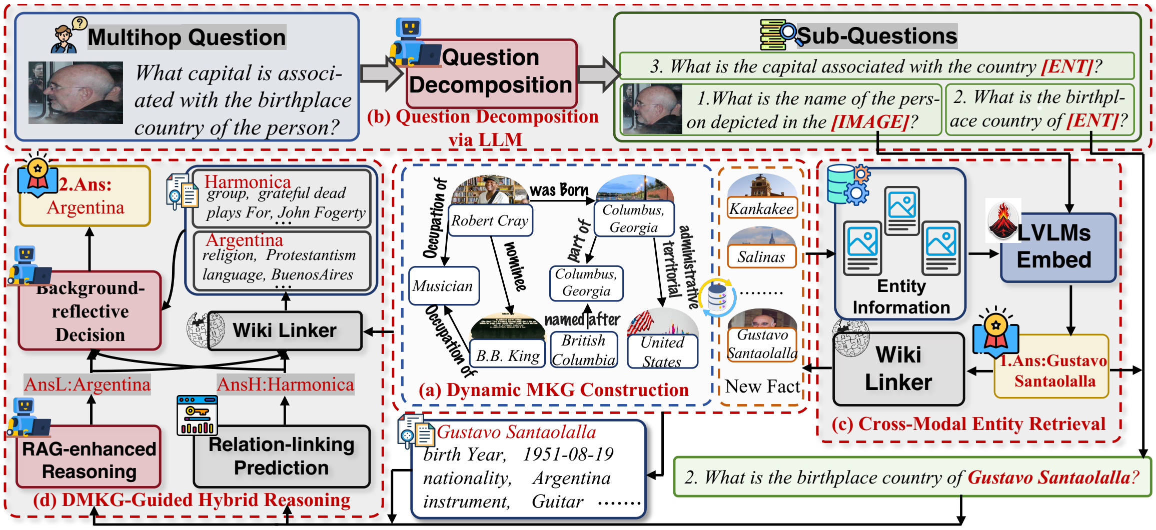


Figure 2: Overall framework of Hybrid-DMKG for MMQAKE task.

ing a clear and editable structure for multihop knowledge updates and multihop traversal. As shown in Figure 2, each statement in $\mathcal{G}$ is represented as $(\mathcal{G}^e_i, \mathcal{G}^r_i, \mathcal{E}^o_i)$, where the head entity $\mathcal{G}^e_i$ and the tail entity $\mathcal{G}^o_i$ belong to the entity set $\mathcal{E}$. Some entity $\mathcal{G}^e_i$ is associated with a corresponding image $\mathcal{G}^v_i$. The relation $\mathcal{G}^r_i$ defines the semantic connection between the head and tail entities.

To incorporate new multimodal knowledge, we integrate an edit quadruple $(x, v, o, \tilde{o})$ into the MKG $\mathcal{G}$, resulting in the DMKG $\tilde{\mathcal{G}}$. The updated MKG $\tilde{\mathcal{G}}$ retains both the original and edited facts, enabling the model to reason over both prior and newly integrated multimodal information.

Question Decomposition Inspired by prior work that decomposes multihop questions into sub-questions to improve retrieval accuracy (Zhong et al. 2023; Gu et al. 2024a; Lu et al. 2025; Li et al. 2025c), we employ an LLM without fine-tuning to decompose multimodal multihop questions. Specifically, we design a template P_{Dec} that transforms a given multimodal multihop question Q into a set of sub-questions:

$$\{q_1, q_2, \ldots, q_n\} = \mathrm{LLM}(Q, P_{\text{Dec}}) \qquad (1)$$

As illustrated by the example in Figure 2, a multihop question is decomposed into three sub-questions. For sub-questions that involve visual information, the placeholder `[IMAGE]` is used to indicate a reference to the image. To promote entity consistency across sub-questions, related entities in other sub-questions are replaced with a special token `[ENT]`. After decomposing the main question, we sequentially address each sub-question.

Cross-Modal Entity Retrieval from DMKG Unlike MQUAKE, which focuses exclusively on textual queries, MMQAKE also requires accurate identification of visual content. For example, answering the sub-question q_1 requires the model to recognize the entity depicted in the rephrased image $\tilde{v}$. Inspired by multimodal retrieval methods (Lewis et al. 2020; Lin et al. 2025), we employ a cross-modal retriever M_u to perform retrieval across different modalities. Specifically, we treat the entities in the DMKG $\tilde{\mathcal{G}}$, each associated with an image and a name, as the candidate answer corpus. To enhance entity representation and retrieval accuracy, we integrate both the image and its linked entity name from $\tilde{\mathcal{G}}$ into a unified representation,

$$z_m = \mathrm{M}_u([\tilde{\mathcal{G}}^e_m, \tilde{\mathcal{G}}^v_m]) \qquad (2)$$

where $\tilde{\mathcal{G}}^e_m$ denotes the entity name of the m-th entity in $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{G}}^v_m$ is its associated image. The sub-question q_1 and the rephrased input image $\tilde{v}$ are encoded using the same module:

$$s = \mathrm{M}_u([q_1, \tilde{v}]) \qquad (3)$$

Then, we identify the most relevant entity a_1 as the subject of next sub-question by computing top1 similarity between query vector s and candidate entity representations z^m:

$$a_1 = \underset{m\in\{1,2,\cdots M\}}{\arg\mathrm{Top1}} \frac{(s)^T z_m}{\|s\|_2\|z_m\|_2} \qquad (4)$$

where M denotes the total number of entities with corresponding images in the DMKG. As shown in Figure 2, the retrieved answer a_1 is *Gustavo Santaolalla*. We replace the `[ENT]` in q_2 with a_1 to form the next-step sub-question: *What is the country of birth of Gustavo Santaolalla?*.

DMKG-Guided Hybrid Reasoning In reasoning sub-questions, such as q_2, we further utilize the DMKG to improve the accuracy and interpretability of answer generation by the LVLM. To retrieve related knowledge from $\tilde{\mathcal{G}}$, we first address variability in natural language expressions (e.g., *United States of America* vs. *USA*, need to refer to the same entity). We apply a **Wiki Linker** [1] module ϕ to normalize the entity a_1, mapping it to its canonical form e_2 within $\tilde{\mathcal{G}}$,

$$e_2 = \phi(a_1) \qquad (5)$$

---

[1] https://wiki.osdev.org/Linker

Input Image	Backbones	Metrics	FT(QFor)	FT(All)	MEND	SERAC	IKE	Ours
Original Image	BLIP-2 (3.8B)	M-Acc	3.73	0.32	0.04	5.75	16.64	**47.55**
		H-Acc	0.20	0.02	0.00	0.00	6.16	**28.88**
	LLaVA (7B)	M-Acc	4.63	1.66	0.70	6.58	38.93	**53.75**
		H-Acc	0.44	0.00	0.00	0.00	16.38	**29.90**
	MiniGPT-4 (7.8B)	M-Acc	4.69	0.08	0.07	0.27	15.48	**35.86**
		H-Acc	0.44	0.00	0.00	0.00	6.14	**24.73**
Rephrased Image	BLIP-2 (3.8B)	M-Acc	0.84	0.04	0.02	1.04	14.44	**45.27**
		H-Acc	0.11	0.00	0.00	0.00	6.06	**26.08**
	LLaVA (7B)	M-Acc	5.71	1.61	0.06	0.97	37.61	**51.27**
		H-Acc	0.53	0.00	0.00	0.02	16.91	**26.16**
	MiniGPT-4 (7.8B)	M-Acc	4.17	0.11	0.04	0.13	9.86	**33.41**
		H-Acc	0.77	0.00	0.00	0.00	5.76	**22.23**

Table 3: Experimental results (%) on the MMQAKE dataset. “QFor” and “All” refer to fine-tuning only the Q-Former parameters and all model parameters, respectively. The best results are highlighted in **bold**.

Based on the linked entity, we extract its associated triples from the $\tilde{\mathcal{G}}$ as a knowledge set,

$$C_2 = \varphi(e_2, \tilde{\mathcal{G}}) \tag{6}$$

where φ denotes the operation that retrieves all relational triples from the DMKG associated with the given entity e_2. The resulting set of associated knowledge is defined as $C_2 = \left\{(e_2, \tilde{\mathcal{G}}^r_{e_2,j}, \tilde{\mathcal{G}}^o_{e_2,j}) \;\middle|\; j = 1, \ldots, k\right\}$. As illustrated in Figure 2, (*birth year*, *1951-08-19*) for the entity *Gustavo Santaolalla*. After obtaining the related knowledge set $\mathcal{C}_2$, we propose a hybrid reasoning module with three parts: (1) *Relation-Link Prediction*, (2) *RAG-Enhanced Reasoning in LVLM*, which together generate candidate answers, and (3) a *Background-Reflective Decision* module that aggregates these candidates to select the most credible response.

(1) Relation-linking Prediction This module leverages explicit DMKG’s relational information for answer prediction. It performs graph-based reasoning over relational paths by assessing the semantic similarity between the query and candidate relation types. Based on our observation, many queries can be answered directly by identifying the underlying relational intent expressed in the question. For example, in query q_2 about **Gustavo Santaolalla**, the implicit relation keyword is “*country of birth*”. If a semantically related relation exists in the DMKG, the most relevant entity can be retrieved and is highly likely to serve as the answer. Motivated by this, we introduce a fine-tuned relation extractor M_e to identify explicit relational keywords k_2^q from the query q_2:

$$k_2^q = \mathrm{M}_e(q_2) \tag{7}$$

The extracted relational keyword k_2^q is then encoded in an embedding $h(k_2^q)$ using a lightweight word embedding *Sense2Vec* (Trask, Michalak, and Liu 2015). Given the candidate answer set C_2 extracted from the knowledge graph, we compute the cosine similarity between the query keyword embedding $h(k_2^q)$ and each candidate relation embedding $h(\tilde{\mathcal{G}}^r_{e_2,j})$. The candidate answer with the highest similarity score is selected as follows:

$$\begin{aligned} j^* &= \arg\max_j \cos\left(h(k_2^q), h(\tilde{\mathcal{G}}^r_{e_2,j})\right) \\ a_2^{\text{link}} &= \begin{cases} \tilde{\mathcal{G}}^o_{e_2,j*}, & \text{if } \cos\left(h(k_2^q), h(\tilde{\mathcal{G}}^r_{e_2,j*})\right) \geq \alpha \\ \varnothing, & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

where j^* denotes the index of the candidate relation that is most semantically aligned with the query. If the similarity score exceeds a threshold α, the corresponding object $\tilde{\mathcal{G}}^o_{e_2,j*}$ is selected as the predicted answer a_2^{link}. Otherwise, if no relevant relation is identified, the answer is indicated by $\varnothing$.

(2) RAG-enhanced Reasoning in LVLM While the linking prediction module is generally effective in identifying target entities and their associated relations, this method may fail when background knowledge is incomplete or when key term extraction is inaccurate. To address this limitation, inspired by the use of retrieval-augmented generation in LLMs (He et al. 2024; Shi et al. 2024) for enhanced reasoning, we propose a RAG-enhanced reasoning module based on DMKG. Specifically, we retrieve the top$K$ knowledge snippets $\mathcal{K}_{\text{Ret}}$ from the associated triple set $\mathcal{C}_2$ that are semantically most relevant to the current query q_2. These retrieved snippets are then incorporated into the answer prompt P_{Ans} and provided as input to the LVLM. This design allows the LVLM to access external knowledge, thereby enhancing its reasoning capabilities when faced with incomplete or ambiguous information,

$$a_2^{\text{model}} = \text{LVLM}\left(q_2, \tilde{v}, \mathcal{K}_{\text{Ret}}\left(q_2, C_2\right), \; P_{\text{Ans}}\right) \tag{9}$$

where a_2^{model} denotes the output of LVLM with RAG. $\mathcal{K}_{\text{Ret}}$ uses the same model architecture as described in Equations (2)–(4), with the key distinction that only the textual modality is employed.

(3) Background-reflective Decision In certain cases, the candidate answers produced by the two reasoning paths differ, i.e., $a_2^{\text{link}} \neq a_2^{\text{model}}$. To resolve such conflicts, we propose a background-reflective decision module. Instead of relying solely on initial predictions, this module enables LVLM to reflectively evaluate competing answers by leveraging the rich semantic and relational context provided by the DMKG. Specifically, for each candidate answer, we extract background information based on the adjacency of the entity level in the DMKG, as determined by the entity link mechanism defined in Equations (5)–(6). The contextual background knowledge representations C_2^{link} and C_2^{modal}, corresponding to a_2^{link} and a_2^{modal}, are defined as follows:

$$\left\{C_2^{\text{link*}}, \; C_2^{\text{modal*}}\right\} = \varphi\Big(\phi\big(\{a_2^{\text{link}}, \; a_2^{\text{modal}}\}\big), \tilde{\mathcal{G}}\Big) \tag{10}$$

These contexts encompass relevant entity descriptions and co-occurring facts that collectively enhance the model's reasoning capabilities. The final prediction is generated by the LVLM using the original question q_2, the associated input image $\tilde{v}$, and answer-related background knowledge as input. This process is formalized as:

$$a_2 = \text{LVLM}(q_2, \tilde{v}, [a_2^{\text{link}}, C_2^{\text{link*}}], [a_2^{\text{modal}}, C_2^{\text{modal*}}], P_{\text{Cho}}) \tag{11}$$

where a_2 denotes the final answer after reflective decision, and P_{Cho} is the choice prompting strategy guiding the reflective reasoning process. In multihop reasoning, we repeat Equations (2)–(4) for sub-questions requiring cross-modal entity retrieval, and apply Equations (5)–(8) during answer generation. By dynamically invoking relevant modules at each reasoning hop, the model gradually resolves multihop questions and generates a final answer.

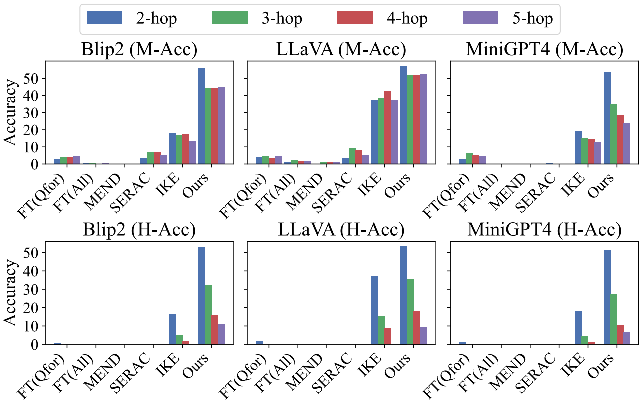
Experiments

We conducted comparative experiments to evaluate the performance of the proposed Hybrid-DMKG method against several existing approaches. The implementation is publicly available at https://github.com/YuanLi95/Hybrid-DMKG.

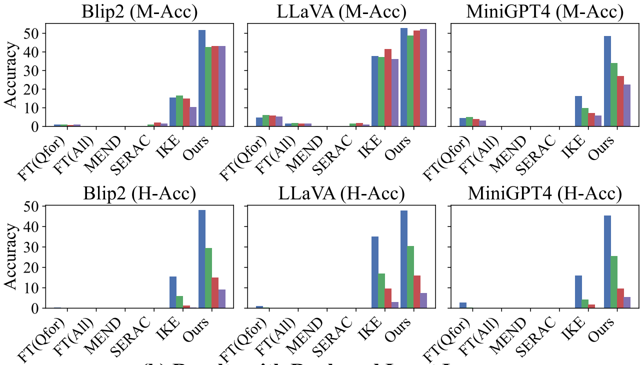
Results

Overall Results Table 3 presents our experimental results on the MMQAKE dataset. With the exception of IKE, most existing MKE methods exhibit significant performance degradation on multihop question answering tasks. Notably, MEND performs the worst, failing to complete any multihop reasoning task, despite demonstrating strong performance on standard single-hop editing tasks (Huang et al. 2024). Moreover, increasing model size does not lead to improved performance. For instance, MiniGPT-4 frequently underperforms compared to the smaller BLIP-2 model. In contrast, IKE, a retrieval-augmented method, maintains relatively stable baseline performance. However, it struggles to integrate multihop information effectively, leading to a significant decline in H-Acc as the number of editing rounds increases.

Our proposed Hybrid-DMKG framework consistently outperforms all baseline methods across various evaluation metrics and backbone configurations. When employing BLIP-2 as the backbone LVLM, Hybrid-DMKG achieves an H-Acc score that surpasses IKE by 22.72% on original images, highlighting its better capability in addressing MMQAKE. Furthermore, with LLaVA as the backbone, Hybrid-DMKG attains M-Acc and H-Acc scores of 53.75% and 29.90%, respectively, demonstrating strong generalizability across different architectures. Notably, the rephrased-image setting poses substantial challenges for multimodal generalization. Under this condition, most models suffer performance degradation, with IKE using MiniGPT-4 experiencing particularly severe declines. Although Hybrid-DMKG also encounters increased difficulty in cross-modal retrieval with rephrased images, it consistently outperforms other baseline models. This robustness mainly stems from the dynamic multimodal knowledge graph, which enhances contextual understanding, as well as the hybrid reasoning framework that performs parallel reasoning and integrates results via a reflective decision module to produce more accurate and reliable outputs.



(a) Results with Original Input Image



(b) Results with Rephased Input Image

Figure 3: Performance comparison of different hops on MMQAKE using the original and rephrased input images.

Results for Different Hops We further analyzed model performance across varying hop counts (2 to 5). As shown in Figure 3, under the M-Acc metric, all models maintain relatively stable performance regardless of hop count. This suggests that multiple valid reasoning paths can lead to correct final answers. However, this stability may also reflect a limitation of the M-Acc metric itself, which only evaluates the final answer correctness and may fail to capture differences in reasoning quality or path validity across varying hop lengths. In contrast, the H-Acc metric imposes a stricter requirement: every intermediate reasoning step must be correct. Consequently, performance consistently declines as the hop count increases. Notably, our model significantly outperforms baselines on 4-hop and 5-hop questions under H-Acc, achieving nearly double their accuracy. In the most challenging 5-hop setting, Hybrid-DMKG exceeds 5% accuracy, while other methods typically remain below 2%. This improvement stems from our effective multihop question decomposition, which mitigates error propagation and reduces hallucinations by large language models. Additionally, the cross-modal retrieval component boosts step-wise accuracy by supplying relevant candidate answers, thereby strengthening the overall reasoning process.

Input Image	Models	BLIP-2		LLaVA		MiniGPT-4	
		M-Acc	H-Acc	M-Acc	H-Acc	M-Acc	H-Acc
Original Image	Ours	47.55	28.88	53.75	29.90	35.86	24.73
	w/o *Linking*	46.09	18.59	47.68	23.15	24.13	14.13
	w/o *RAG*			28.13	21.50		
	w/o *Decision*	48.19	28.05	52.71	28.36	30.44	20.23
Rephrased Image	Ours	45.27	26.08	51.27	26.16	33.41	22.23
	w/o *Linking*	37.22	15.85	43.13	17.95	21.08	8.30
	w/o *RAG*			26.13	19.41		
	w/o *Decision*	44.18	23.76	46.75	23.37	28.53	16.58

Table 4: Ablation study results. The *w/o RAG* setting denotes that only the linking prediction module is used to obtain the answer. As a result, all LVLM backbones yield identical performance under this configuration.

Ablation Study

To assess the contribution of each component in the Hybrid-DMKG framework, we conducted ablation studies on three core modules: Relation-linking Prediction (*Linking*), RAG-enhanced Reasoning in LVLM (*RAG*), and Background-reflective Decision (*Decision*). As shown in Table 4, removing the *Linking* module from MiniGPT-4 leads to a substantially larger performance drop than removing the *RAG* module. For instance, under the rephrased input image setting, the H-Acc score declined by 13.93%. This suggests that MiniGPT-4 struggles to effectively incorporate information from the DMKG, often generating semantically incoherent or irrelevant responses. In contrast, LLaVA demonstrates stronger knowledge aggregation and reasoning abilities, allowing the *RAG* module to operate more effectively and achieve better overall performance. While removing the *Decision* module does not fully disable the model, it causes notable performance drops, especially with rephrased images. This underscores the importance of the background-reflective decision module, which filters incorrect candidates using relevant background knowledge, thereby improving robustness and final answer accuracy.

Related Work

Multimodal Knowledge Editing Methods

Current MKE methods typically adapt existing LLM editing techniques (Touvron et al. 2023) by modifying specific neural network layers and fall into two main categories. (1) **Parameter-update methods** integrate new knowledge into model parameters, such as fine-tuning and MEND (De Cao, Aziz, and Titov 2021), which approximates gradient updates via low-rank decomposition. While effective, these approaches risk catastrophic forgetting, incur high training costs, and can degrade model performance, especially in multihop reasoning (Gu et al. 2024b). (2) **Parameter-retention methods** preserve model parameters and influence outputs through external mechanisms like in-context learning. For example, SERAC (Mitchell et al. 2022) uses a scope classifier and counterfactual modeling, while IKE (Zheng et al. 2023) employs demonstrations to guide edits. However, these methods often depend on task-specific, single-hop textual supervision, limiting their generalization to multihop or cross-modal reasoning. In contrast to existing methods, Hybrid-DMKG uses an MKG to support dynamic knowledge updates and retrieval without modifying model parameters. By integrating cross-modal retrieval, it effectively handles multimodal reasoning over both text and images. In addition, the proposed hybrid reasoning module produces answers through parallel reasoning paths and refines them via reflective decision-making, further improving response accuracy and reliability.

Multihop QA with Knowledge Editing

Recently, to more comprehensively evaluate the reasoning capabilities of KE methods, Zhong et al. (2023) introduced MQUAKE. Unlike traditional KE benchmarks, which primarily assess updates by verifying edited facts or answering single-hop factual queries, MQUAKE emphasizes the model's ability to perform multihop reasoning after knowledge has been injected or updated. This evaluation paradigm is better aligned with the complex reasoning demands typical of real-world (Yuan et al. 2023; Gu et al. 2024a; Shi et al. 2024; Lu et al. 2025). Recent approaches that integrate RAG with question decomposition have demonstrated strong performance in both knowledge editing and reasoning tasks, offering promising directions for advancing the field (Gu et al. 2024a; Shi et al. 2024; Lu et al. 2025; Li et al. 2025b,a).

However, these methods are not directly applicable to MMQAKE, which requires cross-modal knowledge editing and reasoning. To address this, Hybrid-DMKG builds on prior work (Shi et al. 2024; Sun et al. 2024; Sun 2024) with two key enhancements: (1) a cross-modal retrieval model that jointly encodes text and images for accurate entity recognition and multimodal knowledge localization; and (2) a hybrid reasoning module that integrates relation-linking prediction with RAG-enhanced LVLM generation to produce complementary answers. A background-reflective decision module then evaluates these answers using external knowledge, enhancing response consistency and reliability.

Conclusion

In this paper, we introduce MMQAKE, the first benchmark for multimodal multihop question answering with knowledge editing, extending prior multimodal knowledge editing benchmarks. MMQAKE contains questions requiring 2–5 reasoning steps across textual and visual modalities, with an evaluation protocol that verifies factual consistency at each step. To tackle this task, we propose Hybrid-DMKG, a hybrid reasoning framework built on a dynamic multimodal knowledge graph that supports continual knowledge updates. It combines relation-based prediction with RAG using LVLMs to generate parallel answers, and employs a reflective-decision module to improve cross-modal inference and resolve inconsistent reasoning. Experiments show that our method significantly outperforms existing approaches on MMQAKE. Future work will extend MMQAKE with temporal and event-based dynamic knowledge, support open-ended questions beyond factoid QA, and explore end-to-end multihop reasoning without predefined sub-questions.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (62476097, 62276072), the Fundamental Research Funds for the Central Universities, South China University of Technology (x2rjD2250190), Guangdong Provincial Fund for Basic and Applied Basic Research—Regional Joint Fund Project (Key Project) (2023B1515120078), Guangdong Provincial Natural Science Foundation for Outstanding Youth Team Project (2024B1515040010), Guangdong Basic and Applied Basic Research Foundation (2025A1515010162), Guangxi Natural Science Foundation Key Project (No. 2025GXNSFDA069017). TW was supported by the NIHR Maudsley Biomedical Research Centre, Maudsley Charity, King's Together, MHaPS Early Career Researcher Award, and MRC DATAMIND.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.

Chen, K.; Wang, J.; and Zhang, X. 2025. Learning to reason via self-iterative process feedback for small language models. In *Proceedings of the COLING 2025*, 3027–3042.

Chen, S.; Hou, Y.; Cui, Y.; Che, W.; Liu, T.; and Yu, X. 2020. Recall and Learn: Fine-tuning Deep Pretrained Language Models with Less Forgetting. In *Proceedings of the EMNLP 2020*, 7870–7881.

Cui, S.; Zhang, Q.; Ouyang, X.; Chen, R.; Zhang, Z.; Lu, Y.; Wang, H.; Qiu, H.; and Huang, M. 2025. ShieldVLM: Safeguarding the Multimodal Implicit Toxicity via Deliberative Reasoning with LVLMs: ShieldVLM. In *Proceedings of the 33rd ACM MM 2025*, 11677–11686.

De Cao, N.; Aziz, W.; and Titov, I. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the EMNLP 2021*, 6491–6506.

Du, Y.; Jiang, K.; Gao, Z.; Shi, C.; Zheng, Z.; Qi, S.; and Li, Q. 2025. MMKE-Bench: A Multimodal Editing Benchmark for Diverse Visual Knowledge. In *Proceedings of the ICML 2025*.

Gu, H.; Zhou, K.; Han, X.; Liu, N.; Wang, R.; and Wang, X. 2024a. PokeMQA: Programmable knowledge editing for Multi-hop Question Answering. In *Proceedings of the ACL 2024*, 8069–8083.

Gu, J.-C.; Xu, H.-X.; Ma, J.-Y.; Lu, P.; Ling, Z.-H.; Chang, K.-W.; and Peng, N. 2024b. Model Editing Harms General Abilities of Large Language Models: Regularization to the Rescue. In *Proceedings of the EMNLP 2024*, 16801–16819.

He, X.; Tian, Y.; Sun, Y.; Chawla, N.; Laurent, T.; LeCun, Y.; Bresson, X.; and Hooi, B. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Proceedings of the NIPS 2024*, 37: 132876–132907.

Huang, H.; Zhong, H.; Yu, T.; Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2024. VLKEB: a large vision-language model knowledge editing benchmark. In *Proceedings of the NIPS 2024*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proceedings of the NIPS 2020*, 33: 9459–9474.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML 2023*, 19730–19742.

Li, M.; Miao, S.; and Li, P. 2024. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation.

Li, Q.; Li, X.; Chang, Z.; Zhang, Y.; Ji, C.; and Wang, S. 2025a. Multimodal Knowledge Retrieval-Augmented Iterative Alignment for Satellite Commonsense Conversation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 8168–8176.

Li, Q.; Liang, S.; Zhang, Y.; Ji, C.; Chang, Z.; and Wang, S. 2025b. Meta-Knowledge Path Augmentation for Multi-Hop Reasoning on Satellite Commonsense Multi-Modal Knowledge Graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7568–7577.

Li, W.; Wang, J.; Yu, L.-C.; and Zhang, X. 2025c. Topology-of-Question-Decomposition: Enhancing Large Language Models with Information Retrieval for Knowledge-Intensive Tasks. In *Proceedings of the COLING 2025*, 2814–2833.

Liang, D.; Zheng, C.; Wen, Z.; Cai, Y.; Wei, X.-Y.; and Li, Q. 2025a. Seeing Beyond the Scene: Enhancing Vision-Language Models with Interactional Reasoning. *arXiv preprint arXiv:2505.09118*.

Liang, L.; Bo, Z.; Gui, Z.; Zhu, Z.; Zhong, L.; Zhao, P.; Sun, M.; Zhang, Z.; Zhou, J.; Chen, W.; et al. 2025b. Kag: Boosting llms in professional domains via knowledge augmented generation. In *Proceedings of the WWW 2025*, 334–343.

Lin, S.-C.; Lee, C.; Shoeybi, M.; Lin, J.; Catanzaro, B.; and Ping, W. 2025. MM-Embed: Universal Multimodal Retrieval with Multimodal LLMs. In *Proceedings of the ICLR 2025*, 1–20.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Proceedings of the NIPS 2023*, 36: 34892–34916.

Lu, Y.; Zhou, Y.; Li, J.; Wang, Y.; Liu, X.; He, D.; Liu, F.; and Zhang, M. 2025. Knowledge editing with dynamic knowledge graphs for multi-hop question answering. In *Proceedings of the AAAI 2025*, 24741–24749.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022a. Locating and editing factual associations in gpt. *Proceedings of the NIPS 2022*, 35: 17359–17372.

Meng, K.; Sharma, A. S.; Andonian, A. J.; Belinkov, Y.; and Bau, D. 2022b. Mass-Editing Memory in a Transformer. In *Proceedings of the ICLR 2022*.

Mitchell, E.; Lin, C.; Bosselut, A.; Manning, C. D.; and Finn, C. 2022. Memory-based model editing at scale. In *Proceedings of the ICML 2022*, 15817–15831.

Shen, T.; Cambria, E.; Wang, J.; Cai, Y.; and Zhang, X. 2025a. Insight at the right spot: Provide decisive subgraph information to Graph LLM with reinforcement learning. *Information Fusion*, 117: 102860.

Shen, T.; Mao, R.; Wang, J.; Zhang, X.; and Cambria, E. 2025b. Flow-guided Direct Preference Optimization for Knowledge Graph Reasoning with Trees. In *Proceedings of ACM SIGIRR 2025*, 1165–1175.

Shi, Y.; Tan, Q.; Wu, X.; Zhong, S.; Zhou, K.; and Liu, N. 2024. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 2024 ACM CIKM*, 2056–2066.

Sun, X. 2024. *Assessing Model Robustness in Complex Visual Environments*. Ph.D. thesis, The Australian National University (Australia).

Sun, X.; Yao, Y.; Wang, S.; Li, H.; and Zheng, L. 2024. ALICE BENCHMARKS: CONNECTING REAL WORLD RE-IDENTIFICATION WITH THE SYNTHETIC. In *12th International Conference on Learning Representations, ICLR 2024*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trask, A.; Michalak, P.; and Liu, J. 2015. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*.

Wang, J.; Deng, Z.; Deng, D.; Wang, X.; Sheng, R.; Cai, Y.; and Qu, H. 2025a. Empowering multimodal analysis with visualization: A survey. *Comput. Sci. Rev.*, 57: 100748.

Wang, M.; Huang, X.; Xie, J.; Ma, S.; Men, J.; Liang, D.; and Cai, Y. 2025b. From Model Diagram to Code: A Benchmark Dataset and Multi-Agent Framework. In *Proceedings of the ACM MM 2025*, 1754–1763.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the NIPS 2022*, 35: 24824–24837.

Yuan, L.; Cai, Y.; Wang, J.; and Li, Q. 2023. Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. In *Proceedings of the AAAI 2023*, volume 37, 11051–11059.

Zhang, J.; Wang, Z.; Wang, Z.; Zhang, X.; Xu, F.; Lin, Q.; Mao, R.; Cambria, E.; and Liu, J. 2025a. Maps: A multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving. *arXiv preprint arXiv:2503.16905*.

Zhang, J.; Wang, Z.; Zhu, H.; Liu, J.; Lin, Q.; and Cambria, E. 2025b. Mars: A multi-agent framework incorporating socratic guidance for automated prompt optimization. *arXiv preprint arXiv:2503.16874*.

Zhang, J.; Zhang, H.; Yin, X.; Huang, B.; Zhang, X.; Hu, X.; and Wan, X. 2024. Mc-mke: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. *arXiv preprint arXiv:2406.13219*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zheng, C.; Li, L.; Dong, Q.; Fan, Y.; Wu, Z.; Xu, J.; and Chang, B. 2023. Can We Edit Factual Knowledge by In-Context Learning? In *Proceedings of the EMNLP 2023*, 4862–4876.

Zhong, Z.; Wu, Z.; Manning, C. D.; Potts, C.; and Chen, D. 2023. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. In *Proceedings of the EMNLP 2023*, 15686–15702.

Zhu, C.; Rawat, A. S.; Zaheer, M.; Bhojanapalli, S.; Li, D.; Yu, F.; and Kumar, S. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *Proceedings of the ICLR 2024*, 1–17.