MedMKEB: A Comprehensive Knowledge Editing Benchmark for Medical Multimodal Large Language Models

Dexuan Xu[1], Jieyi Wang[2], Zhongyan Chai[2], Yongzhi Cao[1], Hanpin Wang[1], Huamin Zhang[3], Yu Huang[4*]

[1]School of Computer Science, Peking University
[2]School of Software and Microelectronics, Peking University
[3]Institute of Basic Theory of Chinese Medicine, China Academy of Chinese Medical Sciences
[4]National Engineering Research Center For Software Engineering, Peking University

Abstract

Recent advances in multimodal large language models (MLLMs) have significantly improved medical AI, enabling it to unify the understanding of visual and textual information. However, as medical knowledge continues to evolve, it is critical to allow these models to efficiently update outdated or incorrect information without retraining from scratch. Although textual knowledge editing has been widely studied, there is still a lack of systematic benchmarks for multimodal medical knowledge editing involving image and text modalities. To fill this gap, we present MedMKEB, the first comprehensive benchmark designed to evaluate the reliability, generality, locality, portability, and robustness of knowledge editing in medical multimodal large language models. MedMKEB is built on a high-quality medical visual question-answering dataset and enriched with carefully constructed editing tasks, including counterfactual correction, semantic generalization, knowledge transfer, and adversarial robustness. We incorporate human expert validation to ensure the accuracy and reliability of the benchmark. Extensive experiments on state-of-the-art general and medical MLLMs demonstrate the limitations of existing knowledge editing methods in the medical domain, highlighting the need to develop specialized editing strategies.

Datasets — https://github.com/pkusixspace/MedMKEB

Introduction

Medical Multimodal Large Language Models (Medical MLLMs) have become powerful tools with the ability to answer clinical questions, interpret medical images, and support a variety of medical decisions (Xiao et al. 2025). Such models are usually trained on large-scale medical image-text pairs and can capture the complex relationship between vision and language. However, when medical facts change, how to accurately and locally edit existing knowledge in the model without affecting its overall performance remains a key issue that has not been fully explored (Xu et al. 2025).

Knowledge editing is a method that can update, modify, or delete model-specific knowledge without retraining the model. It has become a research hotspot in natural language processing in recent years. Existing work mainly focuses on plain text language models and benchmark datasets, such as ZsRE (Levy et al. 2017) and CounterFact (Meng et al. 2022). However, in the medical field, knowledge is inherently multimodal: medical judgments often rely on the comprehensive analysis of visual evidence, such as radiological images and pathological images, as well as professional text information. Therefore, the multimodal medical knowledge editing task urgently needs new task definitions, benchmark designs, and evaluation methods.

Unlike knowledge editing in general fields, medical knowledge editing is uniquely challenging, and its nature is reflected in high risk, multimodal complexity, and high professionalism (Xu et al. 2024a). First, medical knowledge often relies on the joint understanding of image evidence and text context (Chen et al. 2025b). Second, in clinical applications, incorrect knowledge editing may have serious consequences, so it is not only required to be accurate in facts, but also to comply with the latest medical guidelines and clinical reasoning standards (Chen et al. 2025a). Third, medical concepts usually have a hierarchical structure and are closely related to each other, and knowledge editing requires extremely high precision and granularity (Huang et al. 2025). These characteristics determine that medical knowledge editing requires more professional editing mechanisms and evaluation standards than general methods.

To address these challenges, we propose **MedMKEB**, the first comprehensive benchmark tailored for the task of knowledge editing in medical multimodal large language models. MedMKEB consists of high-quality medical visual question answering data covering multiple medical modalities, tasks, and body parts. It supports systematic evaluation from five fundamental perspectives: **(1) Reliability**: whether the model is consistent with newly injected knowledge; **(2) Locality**: whether irrelevant knowledge is unaffected; **(3) Generality**: whether the model can apply edited knowledge to semantically similar but unseen cases; **(4) Portability**: whether the knowledge can be transferred to related reasoning contexts. In addition, we propose the fifth dimension **(5) Robustness** to test the stability of the edited model under adversarial prompts commonly seen in clinical settings.

In the process of constructing MedMKEB, we designed a series of challenging editing samples, including counterfactual replacement, text semantic rephrasing, image replace-

*Corresponding Author

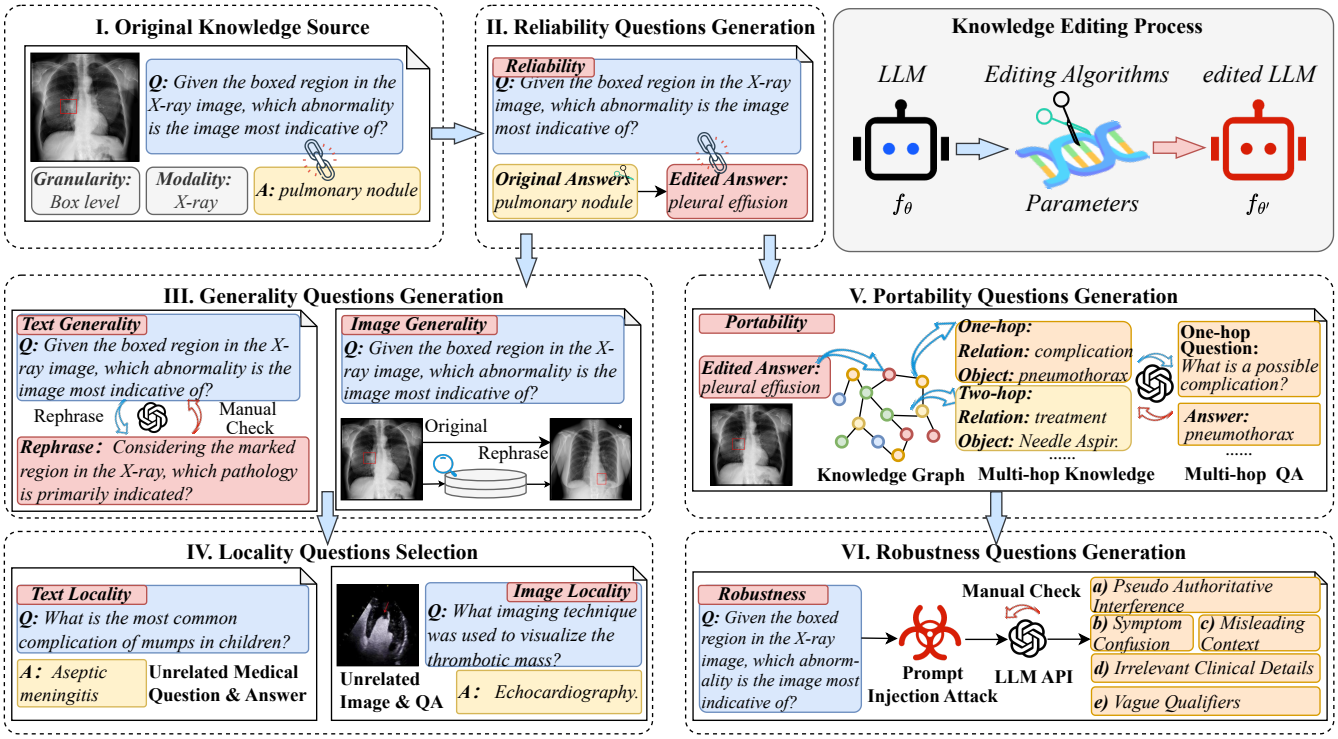

Figure 1: The construction pipeline of MedMKEB.

ment, and multi-hop reasoning chain construction based on a medical knowledge graph. To simulate the interference in the real clinical environment, we also designed a variety of adversarial question injection strategies, such as ambiguous wording, misleading context, redundant clinical information, etc. All generated data are manually reviewed by medical experts to ensure their professionalism and validity. Based on MedMKEB, we test state-of-the-art general and medical multimodal large language models. Extensive experiments have shown that existing knowledge editing methods exhibit limitations in the medical domain, underscoring the need for specialized editing strategies.

In conclusion, our contributions are as follows:

- We propose MedMKEB, the first multimodal medical knowledge editing benchmark, covering diverse modalities, tasks, and expert-verified edits.
- We design a multidimensional evaluation framework that covers five critical aspects, including reliability, locality, generality, portability, and robustness, which provide a holistic view of editing performance in real-world clinical scenarios.
- We conduct large-scale experiments and reveal limitations of existing editing methods, underscoring the need for specialized medical approaches.

Related Work

Knowledge Editing Methods

Knowledge editing has emerged to address issues like knowledge lag, misinformation, and customization needs in Large Language Models (LLMs) by enabling quick and efficient modification or injection of specific knowledge without large-scale retraining. Early methods are mainly based on fine-tuning or parameter-efficient fine-tuning techniques to achieve knowledge modification through local parameter updates. However, these methods are usually costly and have catastrophic forgetting problems. To address these challenges, SERAC (Mitchell et al. 2022b) proposed a knowledge editing method that combines retrieval enhancement and counterfactual comparison, thus leading SERAC to be able to locate contextual information related to the target knowledge and optimize the pertinence and accuracy of editing. IKE (Zheng et al. 2023) emphasizes the realization of knowledge editing through context control without directly modifying model parameters, guiding the model to express updated knowledge during reasoning. Knowledge Editor (De Cao, Aziz, and Titov 2021) avoids catastrophic forgetting caused by the training process by training a hypernetwork. MEND (Mitchell et al. 2022a) proposes to achieve fast and low-cost knowledge editing by learning differentiable gradient transformations, and effectively avoids the impact on other irrelevant knowledge. ROME (Meng et al. 2022) directly modifies model parameters through sparse low-rank matrix updates, injects new factual information into the model, and has stronger controllability.

Knowledge Editing Benchmarks

To assess the effectiveness and reliability of knowledge editing methods, researchers have developed a series of standardized evaluation benchmarks. The commonly used

ZsRE (Levy et al. 2017) dataset tests the accuracy and generative capacity of the model in knowledge editing through sentence-level factual questions and answers. The CounterFact (Meng et al. 2022) benchmark focuses on anti-forgetting and editing scope control, evaluating whether the model retains the original knowledge system while editing new knowledge. In the multimodal field, the benchmark of knowledge editing has made initial progress. MMEdit (Cheng et al. 2023) is the first systematic knowledge editing evaluation framework for MLLMs. The benchmark covers a variety of multimodal tasks such as visual question answering and image caption generation. VLKEB (Huang et al. 2024) constructs further generalization problems by performing multi-hop retrieval in the knowledge graph to test the portability of the model after editing knowledge. MMKE-BENCH (Du et al. 2025) proposes a more challenging multimodal knowledge editing evaluation scheme, including Visual Semantic Editing and User-Specific Editing, which helps promote the practical verification of multimodal knowledge editing methods in complex and real environments. In the medical field, the current main work still focuses on single-modal knowledge editing, such as the counterfactual dataset MedCF (Xu et al. 2024b) and MedEditBench (Chen et al. 2025a), but there is still a lack of knowledge editing evaluation for medical MLLMs. Our work aims to fill this gap.

Preliminary

Problem Definition

Medical multimodal knowledge can be expressed as a triple $k =< i, x, a >$, where i is an image, x is a knowledge question related to the image, and a is the answer to x. In the process of knowledge editing, we hope to modify the answer to obtain $k_e =< i, x, a_e >$, to correct the intrinsic knowledge of the model. Specifically, consider the basic edited dataset: $\mathcal{D}_{edit} = \{(i, x, a, a_e)_j\}_{j=1}^{|\mathcal{D}|}$, each piece of data contains image i, question input x, original answer a, and edited answer a_e. The medical multimodal large language model f_M can be expressed as: $f_M(i, x; \theta) = a$, where $\theta = \theta_v \times \theta_{llm}$, represents the original model parameters, θ_v is the visual module parameter, and θ_{llm} is the language model parameter. After knowledge editing, the model parameters are updated from θ to θ', i.e., $\theta' = \text{KE}(f_M(\theta))$. The expected output of the model for the original input also changes from a to a_e:$f_M(i, x; \theta') = a_e$.

Metrics

In order to comprehensively evaluate the knowledge editing effect of medical multimodal large language models, this benchmark designed a systematic metric system from the dimensions of reliability, locality, generality, and portability (Huang et al. 2024). In addition, we propose the robustness evaluation metric in knowledge editing for the first time. The specific definitions of these metrics are as follows.

Reliability is used to measure the effectiveness of knowledge editing operations on specific goals and ensure that the model output is consistent with the updated medical facts. For a given edit dataset $\mathcal{D}_{edit}$, we sample a quadruple (i, x, a, a_e) from it, and the goal is to modify the original answer a to the expected correct answer a_e. The reliability calculation formula is:

$$\mathcal{M}_{rel} = \mathop{\mathbb{E}}_{(i,x,a,a_e)\sim D_{edit}} [\mathbb{1}\{f_M(i, x; \theta') = a_e\}], \quad (1)$$

where $\mathbb{1}\{\cdot\}$ is an indicator function that returns 1 if the output of the model matches the target answer.

Locality is used to measure the impact of knowledge editing operations on other unedited knowledge and task capabilities of the model. Ideally, editing should be locally effective and have minimal perturbation to global knowledge. Specifically, for the two modalities of image and text, the locality is defined as follows:

$$\mathcal{M}_{loc}^{txt} = \mathop{\mathbb{E}}_{(x,a,a_e)\sim D_{loc}^{txt}} [\mathbb{1}\{f_M(x; \theta') = f_M(x; \theta)\}], \quad (2)$$

$$\mathcal{M}_{loc}^{img} = \mathop{\mathbb{E}}_{(i,x,a,a_e)\sim D_{loc}^{img}} [\mathbb{1}\{f_M(i, x; \theta') = f_M(i, x; \theta)\}], \quad (3)$$

where $\mathcal{D}_{loc}^{txt}$ and $\mathcal{D}_{loc}^{img}$ are the text and visual question-answering dataset outside the distribution of the knowledge editing dataset $\mathcal{D}_{edit}$, respectively.

Generality is used to measure the generalization ability of the model to the new knowledge after editing, that is, whether the model can give answers that are consistent with medical facts in the adjacent semantic space that is not directly edited but related to the target knowledge. This metric reflects that knowledge editing is not limited to single-point corrections. The specific definition is as follows:

$$\mathcal{M}_{gen}^{txt} = \mathop{\mathbb{E}}_{\substack{(i,x,a,a_e)\sim D_{edit} \\ x_g\in\mathcal{N}(x)}} [\mathbb{1}\{f_M(i, x_g; \theta') = a_e\}], \quad (4)$$

$$\mathcal{M}_{gen}^{img} = \mathop{\mathbb{E}}_{\substack{(i,x,a,a_e)\sim D_{edit} \\ i_g\in\mathcal{N}(i)}} [\mathbb{1}\{f_M(i_g, x; \theta') = a_e\}], \quad (5)$$

where $\mathcal{N}(x)$ and $\mathcal{N}(i)$ represent the neighborhood sets of the original input x or i in the semantic space in text and image modalities, respectively.

Portability is used to measure the transferability of knowledge editing capabilities on a wider range of related tasks. It can evaluate whether the edited model can be effectively applied to other knowledge related to the edited knowledge.

$$\mathcal{M}_{port} = \mathop{\mathbb{E}}_{\substack{(i,x,a,a_e)\sim D_{edit} \\ (x_p,a_p)\sim\mathcal{P}(i,x,a,a_e)}} [\mathbb{1}\{f_M(i, x_p; \theta') = a_p\}], \quad (6)$$

where $\mathcal{P}(i, x, a, a_e)$ represents the portability domain related to the original data, such as multi-hop knowledge, relation reversal, task migration, etc.

Robustness is used to measure the stability of the edited knowledge when the model is exposed to adversarial perturbations in prompts, such as misleading rephrasings, distractors, or prompt injections. A robust model should still

Model	Editing Method	Reliability	T-Generality	I-Generality	T-Locality	I-Locality	Portability	Robustness
BLIP2-OPT	FT-LLM	100.0	98.23	99.98	57.63	14.40	24.94	98.73
	FT-Proj	99.09	94.39	99.02	100.0	4.88	27.20	96.26
	IKE	59.12	58.94	59.15	51.11	13.54	22.53	58.44
	SERAC	**99.93**	**99.77**	**99.93**	**100.0**	18.80	24.92	**99.10**
	MEND	99.82	99.60	99.76	99.20	**91.44**	20.27	97.45
	KE	99.19	95.29	99.19	69.05	13.75	**35.15**	93.09
MiniGPT4	FT-LLM	100.0	99.87	100.0	81.22	27.03	44.71	98.97
	FT-Proj	100.0	99.85	99.98	100.0	17.67	50.23	98.22
	IKE	97.39	96.96	97.41	62.51	12.70	51.84	87.32
	SERAC	99.65	99.51	99.65	**99.97**	12.79	57.13	96.09
	MEND	99.91	**99.84**	99.86	99.29	**93.08**	45.69	**97.17**
	KE	**100.0**	99.64	**99.96**	78.71	20.41	**60.05**	96.08
LLaVA	FT-LLM	97.54	95.71	97.52	71.23	18.84	48.01	94.71
	FT-Proj	99.07	96.63	97.02	100.0	12.17	54.35	94.08
	IKE	**100.0**	95.51	**100.0**	71.16	23.02	52.35	92.25
	SERAC	**100.0**	**99.90**	**100.0**	**99.98**	10.19	57.09	**97.83**
	MEND	99.67	99.60	99.69	98.89	**91.77**	46.66	96.79
	KE	99.96	98.23	99.96	82.24	12.30	**62.35**	96.80

Table 1: Single editing results for general MLLMs. The best result and the second best result in each row are indicated by bold and underline, respectively.

output the correct, edited medical knowledge even when the question is subtly altered or perturbed. Given a perturbation function $\mathcal{A}(\cdot)$ that generates adversarial versions of the original input (i, x), the robustness is defined as:

$$\mathcal{M}_{robust} = \mathop{\mathbb{E}}_{\substack{(i,x,a,a_e)\sim D_{edit} \\ (i^{adv},x^{adv})\sim\mathcal{A}(i,x)}} [\mathbb{1}\{f_M(i^{adv}, x^{adv}; \theta') = a_e\}], \tag{7}$$

where (i^{adv}, x^{adv}) denotes the adversarially perturbed image-text pair derived from the original input (i, x), and f_M is expected to preserve the edited knowledge output a_e despite such perturbations.

Benchmark Construction

The overall construction pipeline of MedMKEB is shown in Figure 1, which includes original knowledge source collection, construction of various questions, and manual check.

Original Knowledge Source

Our proposed MedMKEB is built on the GMAI-MMBench (Ye et al. 2024), which provides a comprehensive collection of medical images with relevant clinical questions. GMAI-MMBench is widely used in medical visual question answering, covering multiple medical fields such as radiology, pathology, and ophthalmology. The images, as well as the corresponding questions and answers, provide an ideal data source for developing a knowledge editing benchmark for medical MLLMs.

Specifically, we set rules to filter the modality, task type, department, and perception granularity of questions, avoiding imbalance between categories while ensuring the diversity of questions. In the end, we screened 6987 high-quality multimodal question-answer pairs, covering 16 visual question-answering tasks, 19 human body parts, and 16 types of medical modalities. These question-answer pairs will serve as the MedMKEB's initial knowledge source. The final dataset is denoted as $\mathcal{D}$.

Evaluation Questions Generation

To evaluate the performance of knowledge editing in medical multimodal models, we generate a set of evaluation questions that cover a variety of tasks, including reliability, generality, and locality. These questions are designed to test the model's ability to adapt to edited knowledge while maintaining high performance across different scenarios.

Reliability Questions Generation We construct counterfactual questions that challenge the model to deal with conflicting or newly updated medical knowledge. Counterfactual evaluation is currently used by a large number of knowledge editing benchmarks (Huang et al. 2024; Xu et al. 2024b; Du et al. 2025), and these questions can evaluate whether the model can maintain its accuracy after the information is updated. Specifically, for each visual question-answer pair (i, x, a) in $\mathcal{D}$, we replace the entity a corresponding to the original answer with a different entity with the highest similarity in the options as the editing target a_e, and obtain the updated editing dataset $\mathcal{D}_{edit} = \{(i, x, a, a_e)_j\}_{j=1}^{|\mathcal{D}|}$.

Generality Questions Generation To test the generalization of edited knowledge, we construct new samples that preserve the underlying semantics while varying surface forms. For the textual modality, we apply the LLM API to rephrase the original questions x into semantically equivalent variants x_g, ensuring that the core knowledge remains unchanged. For the visual modality, we identify and randomly replace the original image i with a new image i_g from the dataset that shares the same medical entity. This forms generalized datasets $\mathcal{D}_{gen}^{txt} = \{(i, x_g, a)\}$ and $\mathcal{D}_{gen}^{img} =$

Model	Editing Method	Reliability	T-Generality	I-Generality	T-Locality	I-Locality	Portability	Robustness
Biomed-Qwen2-VL	FT-LLM	100.0	98.37	100.0	45.93	32.37	27.12	99.29
	FT-Proj	100.0	97.73	100.0	1.18	1.01	14.33	95.97
	IKE	<u>99.82</u>	<u>93.11</u>	<u>99.82</u>	60.98	15.69	**38.95**	**90.36**
	SERAC	35.97	34.95	38.02	**99.99**	19.33	19.17	23.03
	MEND	**99.93**	**99.78**	**99.93**	<u>98.98</u>	**92.84**	<u>21.16</u>	<u>89.55</u>
	KE	59.68	55.20	59.77	39.60	<u>29.37</u>	1.32	4.17
LLaVA-Med	FT-LLM	94.53	92.11	90.5	8.58	3.36	15.01	90.8
	FT-Proj	68.74	67.53	64.96	100.0	3.48	10.19	58.58
	IKE	7.86	7.69	7.61	38.36	22.76	1.71	1.48
	SERAC	63.49	57.57	68.85	**100.0**	<u>28.27</u>	8.82	36.02
	MEND	<u>97.59</u>	<u>97.02</u>	<u>97.22</u>	<u>93.67</u>	**56.95**	<u>16.22</u>	<u>94.58</u>
	KE	**99.91**	**98.67**	**98.73**	79.29	2.79	**18.39**	**98.35**
HuatuoGPT-Vision	FT-LLM	100.0	99.02	100.0	34.35	35.78	25.56	98.69
	FT-Proj	93.89	92.94	92.60	97.89	43.03	24.51	73.94
	IKE	<u>99.95</u>	<u>97.68</u>	<u>99.95</u>	60.33	<u>16.68</u>	**42.43**	<u>90.68</u>
	SERAC	57.05	54.32	57.02	**100.0**	11.45	<u>25.95</u>	46.15
	MEND	**99.97**	**99.72**	**99.99**	<u>98.29</u>	**86.99**	23.20	**95.64**

Table 2: Single editing results for medical MLLMs. The best result and the second best result in each row are indicated by bold and underline, respectively.

$\{(i_g, x, a)\}$, where $x_g \in \mathcal{N}(x)$ and $i_g \in \mathcal{N}(i)$, used to evaluate whether the edited knowledge can be transferred and generalized across similar but not identical contexts.

Locality Questions Selection To assess the model's ability to localize and isolate the effects of knowledge editing, we construct locality-sensitive evaluation sets. For the textual modality, we leverage the MedMCQA dataset (Pal, Umapathi, and Sankarasubbu 2022) to identify questions that share similar structures and domains with $\mathcal{D}$ but are not affected by the editing operation. For the visual modality, we use PMC-VQA (Zhang et al. 2023) to sample medically related but semantically independent image-question pairs. This helps verify whether the knowledge editing operation preserves unrelated knowledge and avoids unintended side effects, forming a locality validation set $\mathcal{D}_{loc}^{txt} = \{(x', a')\}$ and $\mathcal{D}_{loc}^{img} = \{(i'', x'', a'')\}$.

Portability and Robustness

We further assess the portability and robustness of edited knowledge in complex medical reasoning and adversarial contexts. These two aspects help to ensure that the modified knowledge is not only effective in isolated cases but can be transferred and defended across more realistic scenarios.

Portability Questions Generation Portability refers to the ability of a model to apply edited knowledge beyond the direct question where editing occurred (Huang et al. 2024), to related contexts within a transferable scope. Formally, we denote this as testing whether the updated knowledge tuple (i, x, a, a_e) can be generalized to a broader reasoning scope $\mathcal{P}(i, x, a, a_e)$. To build the one-hop portability dataset, we identify connected triples of the edited answer entity a_e, such that the model must reason across a single related fact. Suppose the editing operation modifies a sample (i, x, a) into (i, x, a_e), and there exists a fact (a_e, r, o) in the medical knowledge graph. We then construct one-hop reasoning questions $q(o; r)$, which assess whether the model can incorporate a_e in a semantically adjacent question. All such question-answer pairs are collected to form the one-hop portability evaluation dataset $\mathcal{D}_{port}^{(1)}$:

$$\mathcal{D}_{port}^{(1)} = \{q(o; r) | (a_e, r, o) \in \mathcal{K}\}, \tag{8}$$

where $\mathcal{K}$ is the medical knowledge graph and we choose LMKG (Yang et al. 2024) as the reference knowledge graph.

In multi-hop scenarios, evaluating portability becomes more challenging, as the model must integrate the edited knowledge through a chain of reasoning steps. We construct paths $\langle (i, x, a \to a_e), (s_1, r_1, o_1), \ldots, (s_n, r_n, o_n) \rangle$, where $a_e = s_1$ and $o_i = s_{i+1}$ for $i = 1$ to $n - 1$. These paths form a compositional reasoning chain from the editing fact. Similarly, we get the multi-hop knowledge editing dataset:

$$\mathcal{D}_{port}^{(n)} = \{q(o; r_1, r_2, \ldots, r_n) | \mathcal{C}\}, \tag{9}$$

where $\mathcal{C}$ is the multi-hop sequence constructed from the knowledge graph. This design allows us to assess whether the model can propagate the edited knowledge across complex medical reasoning chains, a crucial capability for real-world decision support in clinical settings.

Robustness Questions Generation To test the robustness of knowledge editing, we chose the most common attack method in large language model security, i.e., prompt injection attack (Liu et al. 2023b, 2024; Clusmann et al. 2025). We designed a set of prompt-based perturbation schemes to simulate real attack problems in clinical scenarios (Huang et al. 2025). We apply five types of prompt injection attacks, including *1) pseudo-authoritative interference, 2) symptom confusion, 3) misleading context, 4) irrelevant clinical details, and 5) vague characteristics*. For each question x in $\mathcal{D}_{edit}$, we use the large language model to generate the corresponding adversarial sample x^{adv} and manually verify the generated questions:

$$\mathcal{D}_{adv} = \{(i, x^{adv}, a_e) | x^{adv} \in \mathcal{A}(x)\}, \tag{10}$$

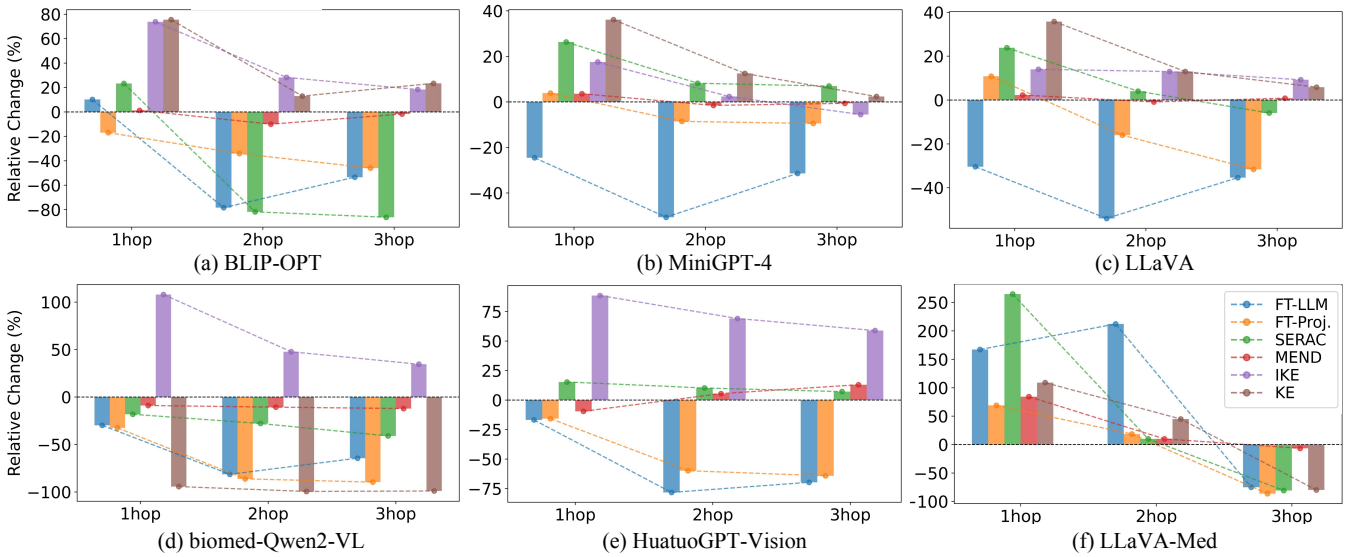

Figure 2: The relative change of multi-hop portability.

where $\mathcal{A}(x)$ is the scope of the perturbed problems generated based on the original question x.

Human Check and Statistics

To ensure the accuracy and consistency of the benchmark, human checks are incorporated into each stage of the construction process. To validate the quality of the answers provided by the model, three medical experts manually review the entire set of questions and correct unreasonable questions. These checks help identify any discrepancies or potential errors in model reasoning and knowledge application.

Statistics MedMKEB contains 6987 knowledge question answering pairs and 13060 images. The knowledge covers 16 medical vqa tasks and includes 2688 portability questions up to 3-hop and 721 robustness questions. We divide the dataset into training and validation sets in a ratio of 6:4.

Experiments

We conducted single and sequential knowledge editing experiments on six MLLMs with different parameter sizes and architectures, including three general models: BLIP-OPT (Li et al. 2023b), MiniGPT-4 (Zhu et al. 2023), and LLaVA (Liu et al. 2023a), and three medical models: LLaVA-Med (Li et al. 2023a), Biomed-Qwen2-VL (Cheng et al. 2024), and HuatuoGPT-Vision (Chen et al. 2024).

Editing Methods and Experimental Settings

Methods Following previous studies, we selected five representative knowledge editing methods as comparison methods, including fine-tuning, Knowledge Editor (KE) (De Cao, Aziz, and Titov 2021), MEND (Mitchell et al. 2022a), SERAC (Mitchell et al. 2022b), and IKE (Zheng et al. 2023). Fine-tuning includes fine-tuning of the LLMs (FT-LLM) or the visual language alignment module (FT-Proj). To avoid catastrophic forgetting, we only fine-tune the last layer of the large language model.

Settings We use the EasyEdit framework (Wang et al. 2024) and complete the support for medical multimodal large language models. All experiments are performed on one NVIDIA A800 GPU with 80GB. For each model, we use its public pre-trained weights. The batch size of all experiments is uniformly set to 1, the text editing loss weight and image editing loss weight are set to 0.1, and the locality loss weight is set to 1.

Single Editing Results and Analysis

We first conducted single editing experiments. Tables 1 and 2 show the results for general models and medical models, respectively. From the observations, we can draw the following conclusions. 1) For general models, the reliability of various algorithms is very close, and in most cases it can reach more than 99%, indicating that these algorithms can successfully modify the edited knowledge. For medical models, the reliability of the SERAC algorithm decreased significantly, and none of them reached 70%, which is related to the ability of the counterfactual model. 2) Compared to the fine-tuning method, on different models, several carefully designed knowledge editing algorithms can have higher locality while ensuring generalization. It shows that knowledge editing algorithms can effectively avoid the influence of irrelevant knowledge. The MEND algorithm has a higher I-Locality score than other algorithms, whether it is a general model or a medical model. 3) For general models and LLaVA-Med, KE obtained the best portability results. This shows that hypernetwork-based editing methods can more effectively cope with this challenge. In addition, all knowledge editing methods have good generalization, but have difficulties in portability. This highlights the difficulty in applying edited knowledge to new content. 4) Regarding

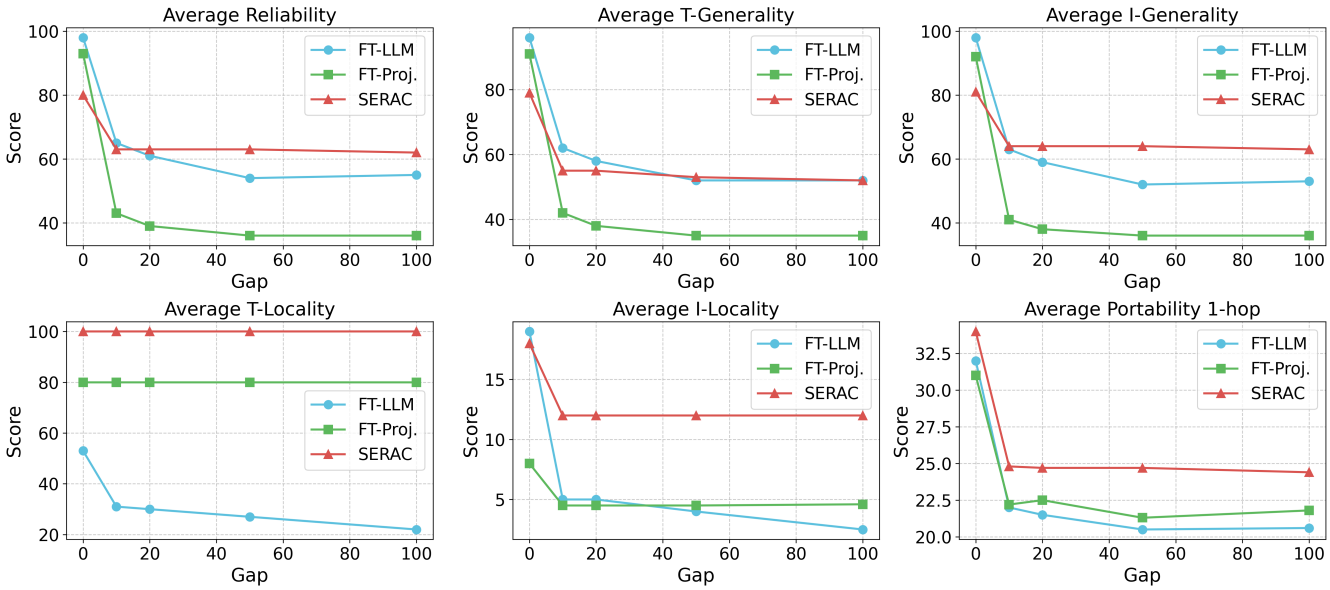

Figure 3: Average results in sequential editing.

robustness, existing knowledge editing methods show a decline in both general models and medical models. However, FT-LLM hardly shows a decline in robustness, which indicates that the existing knowledge editing algorithms lack defense against prompt injection attacks.

Existing medical model knowledge editing algorithms still need significant improvement. 1) Model aspect. Medical MLLMs are often obtained by direct fine-tuning from general MLLMs, which overfit to the training data, resulting in reduced generality when facing unseen data. At the same time, the model's context learning ability is weakened, and knowledge editing algorithms such as IKE and SERAC, whose parameters are not updated, cannot answer questions correctly. 2) Algorithm aspect. The existing knowledge editing algorithms for MLLMs mainly focus on LLM parameters and lack joint optimization of visual modules and text modules, resulting in poor performance of the algorithm's image locality evaluation indicators. Among these algorithms, meta-learning-based algorithms such as KE and MEND perform well, but are still not as good as those in general models. These conclusions show that there is still a need to design specific knowledge editing algorithms for medical MLLMs.

Multi-hop Portability Results

To comprehensively evaluate portability, we show the relative change in portability performance over multiple hops in Figure 2. In general, the relative performance of almost all models and methods shows a downward trend with the increase of the number of hops, indicating that knowledge transfer faces greater challenges in multi-hop reasoning paths. In terms of model dimension, LLaVA-Med performs best in the 1-hop scenario, but this advantage rapidly weakens as the number of hops increases, and even turns to negative growth in the 3-hop scenario, indicating that the model is extremely sensitive to local knowledge updates but unstable in long-distance transmission. The IKE method shows strong stability in multiple models, and its performance degradation is significantly smaller than that of other methods. In addition, SERAC and MEND show good anti-degradation capabilities in some models, especially in the 2-hop and 3-hop scenarios of MiniGPT-4 and HuatuoGPT-Vision, which can still maintain relatively stable performance. The FT-LLM method shows drastic fluctuations in multi-hop tasks, indicating that it is difficult to support the deep transmission of complex knowledge by only adjusting the parameters of the last layer of LLM.

Sequential Editing Results and Analysis

We also conducted experiments on sequential editing, and the average results of the 3 algorithms are shown in Figure 3. We set the sequence editing gaps to 10, 20, 50, and 100. The experimental results show that SERAC shows good stability in all metrics, especially when dealing with large gaps. As the gap increases, the reliability of FT decreases, and the locality decreases significantly. This shows that the effect of sequential editing needs to be improved.

Conclusion

In this paper, we propose MedMKEB, the first benchmark for evaluating knowledge editing in multimodal large-scale language models for medicine. MedMKEB evaluates five key aspects: reliability, locality, generality, portability, and robustness. Extensive experiments show that current editing methods struggle to handle the high precision and multimodal nature of medical knowledge. MedMKEB fills a critical gap and lays the foundation for developing safer and more effective knowledge editing techniques in medicine.

Acknowledgments

This paper was supported by National Key R&D Program of China (No. 2023YFC3502902), National Natural Science Foundation of China under Grants (62436006, 62172016), Sanya Science and Technology Special Fund (No. 2024KFJX04), Beijing Natural Science Foundation (No. L257018) and Beijing Nova Program.

References

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; et al. 2024. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Chen, S.; Luo, L.; Qiu, Z.; Cao, Y.; Yang, C.; and Pan, S. 2025a. Beyond Memorization: A Rigorous Evaluation Framework for Medical Knowledge Editing. *arXiv preprint arXiv:2506.03490*.

Chen, Y.; Xu, D.; Huang, Y.; Zhan, S.; Wang, H.; Chen, D.; Wang, X.; Qiu, M.; and Li, H. 2025b. MIMO: A Medical Vision Language Model with Visual Referring Multimodal Input and Pixel Grounding Multimodal Output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24732–24741.

Cheng, D.; Huang, S.; Zhu, Z.; Zhang, X.; Zhao, W. X.; Luan, Z.; Dai, B.; and Zhang, Z. 2024. On Domain-Specific Post-Training for Multimodal Large Language Models. *arXiv preprint arXiv:2411.19930*.

Cheng, S.; Tian, B.; Liu, Q.; Chen, X.; Wang, Y.; Chen, H.; and Zhang, N. 2023. Can We Edit Multimodal Large Language Models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13877–13888.

Clusmann, J.; Ferber, D.; Wiest, I. C.; Schneider, C. V.; Brinker, T. J.; Foersch, S.; Truhn, D.; and Kather, J. N. 2025. Prompt injection attacks on vision language models in oncology. *Nature Communications*, 16(1): 1239.

De Cao, N.; Aziz, W.; and Titov, I. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 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 *The Thirteenth International Conference on Learning Representations*.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Huang, X.; Wang, X.; Zhang, H.; Zhu, Y.; Xi, J.; An, J.; Wang, H.; Liang, H.; and Pan, C. 2025. Medical mllm is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3797–3805.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-shot relation extraction via reading comprehension. In *21st Conference on Computational Natural Language Learning, CoNLL 2017*, 333–342. Association for Computational Linguistics (ACL).

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, X.; Yu, Z.; Zhang, Y.; Zhang, N.; and Xiao, C. 2024. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*.

Liu, Y.; Deng, G.; Li, Y.; Wang, K.; Wang, Z.; Wang, X.; Zhang, T.; Liu, Y.; Wang, H.; Zheng, Y.; et al. 2023b. Prompt Injection attack against LLM-integrated Applications. *arXiv preprint arXiv:2306.05499*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.

Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2022a. Fast Model Editing at Scale. In *International Conference on Learning Representations*.

Mitchell, E.; Lin, C.; Bosselut, A.; Manning, C. D.; and Finn, C. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, 15817–15831. PMLR.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Wang, P.; Zhang, N.; Tian, B.; Xi, Z.; Yao, Y.; Xu, Z.; Wang, M.; Mao, S.; Wang, X.; Cheng, S.; et al. 2024. EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 82–93.

Xiao, H.; Zhou, F.; Liu, X.; Liu, T.; Li, Z.; Liu, X.; and Huang, X. 2025. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117: 102888.

Xu, D.; Chen, Y.; Chai, Z.; Xiao, Y.; Yan, Y.; Ding, W.; Wang, H.; Jin, Z.; Jiao, W.; Yue, W.; et al. 2025. Knowledge fusion in deep learning-based medical vision-language models: A review. *Information Fusion*, 103455.

Xu, D.; Chen, Y.; Wang, J.; Huang, Y.; Wang, H.; Jin, Z.; Wang, H.; Yue, W.; He, J.; Li, H.; et al. 2024a. Mlevlm:

Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, 4977–4997.

Xu, D.; Zhang, Z.; Zhu, Z.; Lin, Z.; Liu, Q.; Wu, X.; Xu, T.; Wang, W.; Ye, Y.; Zhao, X.; et al. 2024b. Editing factual knowledge and explanatory ability of medical large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2660–2670.

Yang, P.; Wang, H.; Huang, Y.; Yang, S.; Zhang, Y.; Huang, L.; Zhang, Y.; Wang, G.; Yang, S.; He, L.; et al. 2024. LMKG: A large-scale and multi-source medical knowledge graph for intelligent medicine applications. *Knowledge-Based Systems*, 284: 111323.

Ye, J.; Wang, G.; Li, Y.; Deng, Z.; Li, W.; Li, T.; Duan, H.; Huang, Z.; Su, Y.; Wang, B.; et al. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

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 2023 Conference on Empirical Methods in Natural Language Processing*, 4862–4876.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *The Twelfth International Conference on Learning Representations*.