MultiMedBench: A Scenario-Aware Benchmark for Evaluating Knowledge Editing in Medical VQA

Shengtao Wen[1*], Haodong Chen[1*], Yadong Wang[1], Zhongying Pan[2], Xiang Chen[1†], Yu Tian[3], Bo Qian[1], Dong Liang[1], Sheng-Jun Huang[1],

[1]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]Huaneng Information Technology Co., Ltd.
[3]Tsinghua University
{shengtao_wen, xiang_chen}@nuaa.edu.cn

Abstract

Knowledge editing (KE) provides a scalable approach for updating factual knowledge in large language models without full retraining. While previous studies have demonstrated effectiveness in general domains and medical QA tasks, little attention has been paid to KE in multimodal medical scenarios. Unlike text-only settings, medical KE demands integrating updated knowledge with visual reasoning to support safe and interpretable clinical decisions. To address this gap, we propose **MultiMedBench**, the first benchmark tailored to evaluating KE in clinical multimodal tasks. Our framework spans both *understanding* and *reasoning* task types, defines a three-dimensional metric suite (reliability, generality, and locality), and supports cross-paradigm comparisons across general and domain-specific models. We conduct extensive experiments under single-editing and lifelong-editing settings. Results suggest that current methods struggle with generalization and long-tail reasoning, particularly in complex clinical workflows. We further present an efficiency analysis (e.g., edit latency, memory footprint), revealing practical trade-offs in real-world deployment across KE paradigms. Overall, MultiMedBench not only reveals the limitations of current approaches but also provides a solid foundation for developing clinically robust knowledge editing techniques in the future.

Code — https://github.com/NUAA-MMMI/MedBench

Introduction

Multimodal Large Language Models (MLLMs) have witnessed rapid advancements, with breakthroughs in visual-language alignment and scalable model architectures enabling strong performance across a range of general tasks. However, applying these models effectively in high-stakes domains such as medicine remains an open challenge. Despite initial success in tasks like medical image interpretation, clinical question answering, and decision support, solely relying on model scaling or domain adaptation remains far from sufficient. As illustrated in Figure 1, general-purpose MLLMs still face structural limitations in clinical settings—including outdated knowledge (Wu et al. 2025; Kim et al. 2025), high heterogeneity across modalities (Chen et al. 2024c; Dai et al. 2025), and strict demands for safety, interpretability, and compliance (Wang et al. 2024b; Han et al. 2024; Chen et al. 2022b).

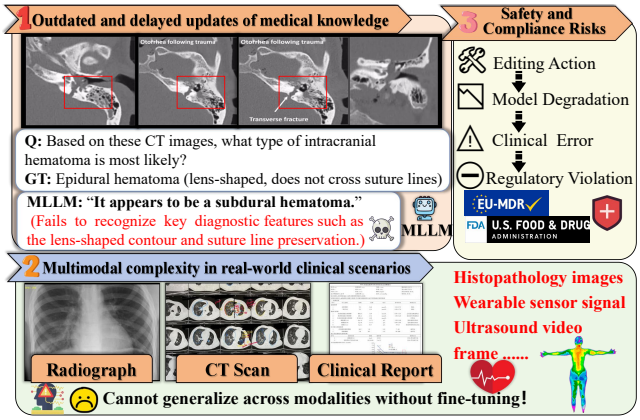


Figure 1: Key challenges faced by general-purpose MLLMs in clinical applications: (1) outdated medical knowledge can lead to inaccurate or unsafe outputs; (2) high diversity in data modalities and task types demands robust multimodal understanding; (3) safety-critical requirements necessitate traceable and reliable model behavior.

Real-world medical practice evolves continuously with the approval of new therapies, revisions of treatment guidelines, and discoveries from ongoing clinical trials. In such a dynamic and rapidly changing landscape, deploying frozen MLLMs trained on static corpora creates a widening gap between model behavior and clinical reality. This highlights a crucial but underexplored challenge: how can we efficiently and safely inject new medical knowledge into existing MLLMs without compromising prior capabilities? Traditional fine-tuning offers limited solutions to these challenges (Luo et al. 2023). It typically requires full-network

*These authors contributed equally.
†Corresponding author.

optimization, incurs high computational costs, and involves long iteration cycles. Furthermore, it is prone to catastrophic forgetting, where new knowledge overwrites existing competencies and severely degrades performance on previously learned tasks (Kalajdzievski 2024; Chen et al. 2022c).

Knowledge Editing (KE) has emerged as a more surgical and efficient alternative to traditional fine-tuning. Instead of retraining the entire model, KE introduces localized updates—via selective parameter modification or external memory injection—allowing models to incorporate new facts rapidly while preserving global behavior (Mitchell et al. 2021; Yao et al. 2025). Its efficiency, low interference, and auditability make it especially suited for safety-critical domains like healthcare (Youssef et al. 2025).

Despite recent advances in KE for general-purpose language models, its applicability to clinical domains remains largely underexplored. Existing benchmarks, such as MedEditBench (Xu et al. 2024b), focus primarily on textual edits and often overlook challenges fundamental to real-world healthcare, including multimodal fusion, clinical reasoning, among others. To truly assess and enhance the potential of models in real-world healthcare settings, the field urgently requires a new, multimodally-centered benchmark.

To address this need, we present MultiMedBench, a benchmark specifically designed for evaluating knowledge editing in clinical multimodal contexts. Unlike existing knowledge-editing benchmarks such as MedEditBench, which are limited to textual QA and static fact correction. MultiMedBench is the first to support *multimodal* and *VQA-based* evaluation in medical settings. This introduces unique challenges: models must not only incorporate new knowledge, but also interpret complex image-text inputs, localize lesions, reason across time, and adapt to diverse clinical modalities. These requirements significantly raise the bar for reliable and safe knowledge editing in real-world clinical scenarios. The benchmark encompasses the following components: (1) **Dual-Axis Task Design**: Tasks are structured along two dimensions, including task type (*understanding* vs. *reasoning*) and input modality (text plus single-frame or multi-frame images), covering the full spectrum from visual recognition to multimodal diagnostic inference. (2) **Three-Dimensional Evaluation Metrics**: We propose a comprehensive framework consisting of reliability (accuracy on edited targets), generality (robustness to semantic variations), and locality (preservation of unrelated outputs), enabling systematic analysis of both effectiveness and side effects. (3) **Cross-Paradigm Method Comparison**: We evaluate four representative KE paradigms, namely Prompt, LoRA (Zhang et al. 2023b), GRACE (Hartvigsen et al. 2023), and WISE (Wang et al. 2024a), under both single and lifelong editing settings across general-purpose MLLMs and domain-specific medical models.

Extensive experiments yield three key findings. First, existing KE methods underperform on complex long-tail reasoning tasks in medical contexts. Second, lifelong editing introduces order dependence and catastrophic forgetting, reducing model stability. Third, most methods are limited to short-text or atomic fact edits and cannot support the depth and contextual richness required in realistic clinical scenarios. In general, our contributions are threefold:

- We propose the first comprehensive benchmark specifically designed for multimodal medical knowledge editing, targeting critical challenges in clinical AI evaluation.
- A unified three-dimensional evaluation framework is established to quantify editing effectiveness, generalization, and unintended side effects in medical settings.
- Extensive empirical analysis highlights the critical limitations of current KE methods in handling complex clinical reasoning, thereby offering a foundation for future research and method development.

We envision MultiMedBench as a crucial stepping stone toward clinically reliable knowledge updating protocols in future medical foundation models.

Related Work

Knowledge Editing for LLMs

Knowledge Editing seeks to enable accurate and semantically coherent responses from large language models (LLMs) by selectively updating internal knowledge representations without resorting to full model retraining (Yao et al. 2023). As LLMs scale and are increasingly deployed in real-world applications, ensuring the timeliness and factual consistency of their embedded knowledge becomes essential, particularly in high-stakes fields like healthcare.

Previous studies have shown that most LLMs struggle to adapt to time-sensitive updates unless explicitly augmented or edited (Li et al. 2024; Wu et al. 2024; Dhingra et al. 2022). To address this, research has proposed various knowledge editing paradigms (Zhang et al. 2024), broadly categorized as: (1) **External Retrieval-Based Approaches**, which avoid modifying model weights by retrieving relevant facts from external memory or tools (Zheng et al. 2023; Mitchell et al. 2022b; Jiang et al. 2024); (2) **Latent State Injection**, which integrates new knowledge by altering internal representations (Hartvigsen et al. 2023; Yu et al. 2024; Hu et al. 2022; Dettmers et al. 2023; Zhang et al. 2023c); and (3) **Internal Structure Editing**, which directly edits model parameters to encode persistent knowledge updates (Meng et al. 2022, 2023; Mitchell et al. 2022a; Fang et al. 2024; Cai, Cao, and et al. 2024; Feng et al. 2025).

Toward Knowledge-Adaptive MLLMs in Medicine

Multimodal large language models (MLLMs) such as Flamingo (Alayrac et al. 2022), PaLI (Chen et al. 2022a), and GPT-4V (Yang et al. 2023) have significantly advanced vision-language tasks, including VQA, image captioning, and multimodal instruction following. Extending this progress to the medical domain, models like HuatuoGPT-Vision (Chen et al. 2024a), LLaVA-Med (Li et al. 2023), and Med-Flamingo (Boger et al. 2023) demonstrate that domain-specific fine-tuning on radiological images and clinical reports improves grounding and interpretability. XrayGPT (Zhang et al. 2023a) and ChatRad (Huang et al. 2024b) further explore diagnostic reasoning and radiology-focused VQA, indicating early promise in real-world clinical tasks.

However, deploying MLLMs in clinical settings remains challenging. Medical applications require models to maintain up-to-date medical knowledge, yet most existing systems are trained statically and lack mechanisms for post-deployment updates. This can lead to hallucinations and outdated recommendations (Yan et al. 2024; Chen et al. 2024d). Knowledge editing offers a solution by enabling targeted model updates without full retraining. Early efforts like MedLaSA (Xu et al. 2024a) adopt adapter-based strategies to improve factual accuracy in medical LLMs. Nonetheless, most prior work remains unimodal and lacks support for editing visual knowledge. Furthermore, current benchmarks primarily evaluate general-domain or text-only edits (He 2024; Hu et al. 2024), leaving an important gap in evaluating multimodal, domain-specific model adaptation. To bridge this gap, we introduce the first benchmark dataset for multimodal knowledge editing in medical MLLMs.

Benchmark Construction

Inspired by recent knowledge editing applications in unimodal settings, the constructed MultiMedBench dataset is composed entirely of visual question answering data. The construction process is illustrated in Figure 3.

Preliminaries

Knowledge editing aims to adjust the behavior of a base model f_θ (where θ denotes the model parameters) with respect to a edit descriptor (x_e, y_e), while preserving the model's performance on other samples. The ultimate goal is to obtain an edited model, denoted as f_{θ_e} (Yao et al. 2023).

The base model f_θ is defined as a function $f : X \to Y$ that maps an input x to a predicted output y. Given an edit input x_e and target label y_e such that $f_\theta(x_e) = y_e$, the edited model is expected to satisfy $f_{\theta_e}(x_e) = y_e$.

Knowledge editing typically affects a set of inputs closely associated with the edit example, referred to as the *editing scope*. A successful edit should modify the model's behavior within the scope, while keeping predictions unchanged for out-of-scope inputs. Formally:

$$f_{\theta_e}(x) = \begin{cases} y_e, & \text{if } x \in I(x_e, y_e) \\ f_\theta(x), & \text{if } x \in O(x_e, y_e) \end{cases} \quad (1)$$

Here, $I(x_e, y_e)$ denotes in-scope inputs, typically including x_e and its equivalence neighborhood $N(x_e, y_e)$; whereas $O(x_e, y_e)$ refers to inputs unrelated to the edit descriptor.

Design Principle

To systematically evaluate medical multimodal knowledge-editing methods, we adopt a two-tier experimental taxonomy: the *Understanding* tier requires models to integrate medical images with clinical narratives to deliver a coherent explanation of the patient's condition, whereas the *Reasoning* tier demands cross-view and temporal inference over multi-frame studies to support complex clinical decisions. This hierarchical design assesses model competence from basic visual recognition to full multimodal diagnostic reasoning. The benchmark encompasses two input modalities: *single-frame* images and *multi-frame* images. Multiple metrics are designed to evaluate knowledge editing methods, such as reliability, generality and locality.

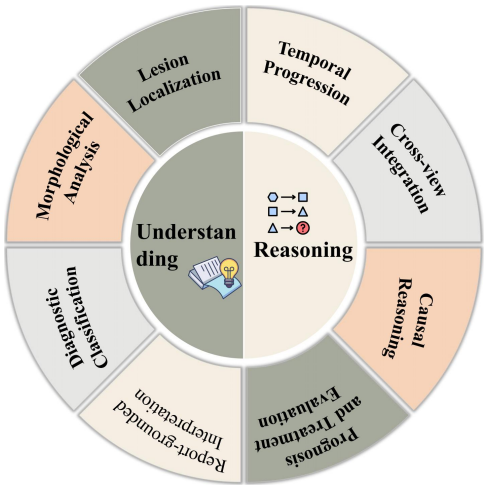


Figure 2: The statistics of scenario types for MultiMedBench, encompassing two principal categories of clinical tasks: *understanding* and *reasoning*.

Scenario Type. The *Understanding* scenario requires models to fuse one or a few medical images with accompanying clinical narratives—such as chief complaints, radiological findings, and medical history—to produce a semantically consistent explanation covers lesion location, characteristics, staging, and potential management strategies. The *Reasoning* scenario represents a more demanding tier: models perform cross-view and temporal inference over multi-frame or multi-view studies, tracking lesion dynamics and jointly leveraging imaging, textual, and temporal cues to support disease-course analysis, treatment-response evaluation, and prognostic prediction. Relative to the *Understanding* tier, the *Reasoning* tier places greater emphasis on temporal modelling, cross-view alignment, and causal reasoning. Statistics for each data type are summarised in Table 1. The distribution of scenario types is provided in Figure 2.

Data Type. The *Single-frame* modality comprises a single static CT or MRI slice, or an ultrasound frame, accompanied by a concise clinical narrative. It targets foundational visual perception and local diagnostic tasks, such as lesion localisation and morphological assessment. The *Multi-frame* modality consists of time-series or multi-view images acquired from the same anatomical region, together with aligned textual descriptions, and specifically is intended to probe a model's ability in temporal modelling, cross-view fusion, and dynamic lesion analysis.

Metrics. *Reliability* denotes the post-editing hit rate on the target samples, reflecting whether the intended knowledge has been correctly injected. *Generality* measures the proportion of correct responses under semantically equivalent paraphrases, gauging the transferability of the edit. *Locality* quantifies the extent to which predictions on unrelated tasks

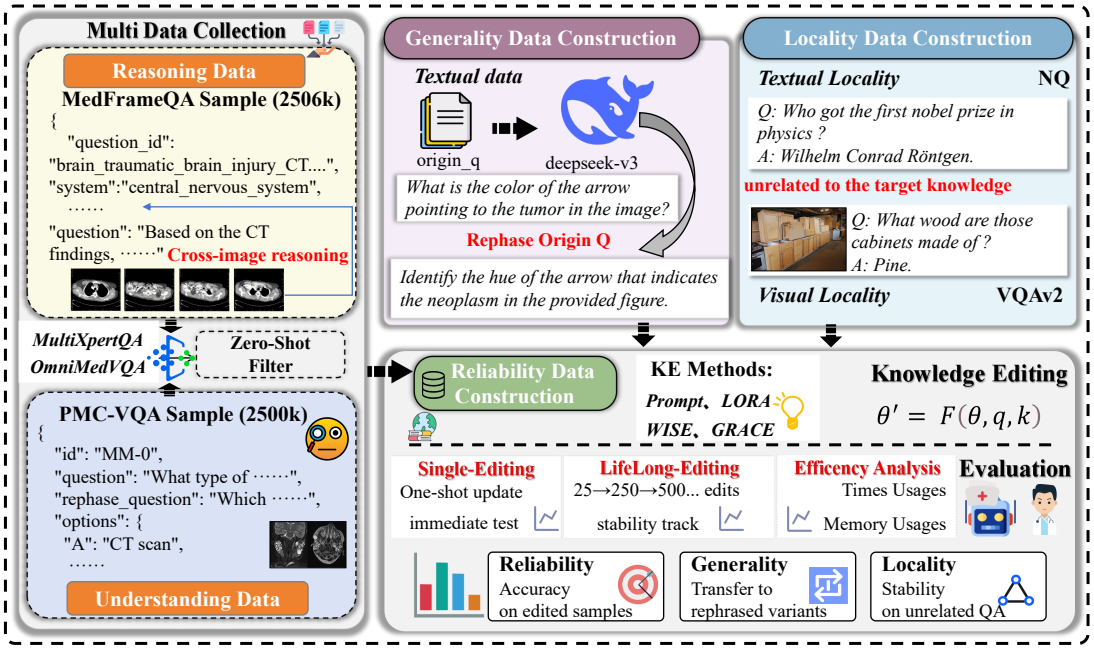


Figure 3: Overview of the MultiMedBench pipeline. Evaluation subsets (reliability, generality, locality) are built via zero-shot filtering and external sources (DeepSeek-V3, VQAv2, NQ). Four editing methods (Prompt, LoRA, GRACE, WISE) are tested on two general MLLMs—LLaVA-OneVision and Qwen2-VL—and one medical MLLM, HuatuoGPT, across multiple criteria.

or samples remain unchanged after editing, thereby assessing the edit's side-effect footprint.

Data Collection

To construct a representative, challenging, and evaluation–friendly benchmark for multimodal medical knowledge editing, we perform a systematic filtering and hierarchical restructuring of publicly available resources. The resulting dataset comprises multiple high-quality subsets tailored to two core tasks: *Understanding* and *Reasoning*.

Data Sources. We construct our comprehensive medical multimodal knowledge editing benchmark based on three high-quality public datasets: MedFrameVQA (Liu et al. 2024) for reasoning-oriented tasks, PMC-VQA (He et al. 2023) for understanding-oriented tasks, and selected samples from both MedXpertQA (Cheng et al. 2024) and OmniMedVQA (Huang et al. 2024a) to enhance task diversity and overall knowledge coverage.

To ensure the benchmark's difficulty and diagnostic value, we employ a *zero-shot filtering* strategy. We use medical MLLM `Radiology-Infer-Mini` to perform inference over all raw samples and retain only those for which the model fails to produce correct answers. This results in two challenging subsets, $MultiMedBench_U$ and $MultiMedBench_R$, designed to focus evaluation on model deficiencies and increase the signal-to-noise ratio.

Construction of the Reliability Subset. The filtered samples from $MultiMedBench_U$ and $MultiMedBench_R$ together constitute the Reliability subset, which is specifically used to assess whether injected knowledge is correctly reflected in the model's updated responses.

Subset	Understanding	Reasoning	Total
Training	3161	3257	6418
Testing	1355	1396	2751

Table 1: Number of samples by task type and modality in MultiMedBench

Construction of the Generality Subset. To evaluate a model's generalization ability under linguistic variation, we construct the Generality-Text subset. Using the DeepSeek-v3 model, we generate modified versions of the original questions that preserve semantic meaning and ground-truth answers, resulting in semantically equivalent question pairs with various surface forms. Details on prompt templates and rewriting heuristics are provided in the Appendix D.

Construction of the Locality Subset. To assess the potential side effects of knowledge editing on seemingly unrelated tasks, we construct two locality-focused evaluation subsets. The first, Locality-Text, is randomly sampled from the Natural Questions dataset (Kwiatkowski et al. 2019) to quantify whether the edited model still maintains its performance on general-domain language understanding tasks. The second, Locality-Modality, is based on VQAv2 (Antol et al. 2015) and includes open-domain natural image question-answer pairs to measure whether the model successfully retains general visual reasoning capabilities in non-

medical contexts after medical knowledge editing.

Quality Control. To enhance the dataset's validity and robustness, we have implemented a multi-stage quality control pipeline. For consistency and ambiguity, three graduate-level annotators have reviewed a random 20% sample of paraphrased and generalized question-answer pairs to assess semantic equivalence and clarity, and have resolved disagreements by majority vote. To validate this review process, we have measured inter-annotator consistency across 200 random samples, which has revealed a disagreement rate of less than 7%. These discrepancies have been settled through majority voting and adjudicated revision.

Editing Method Selection

To broadly compare intrinsic knowledge editing paradigms, we select representative methods from three major categories. LoRA represents the fine-tuning paradigm, using low-rank adapters to efficiently update a subset of weights. WISE exemplifies memory-enhanced editing via addressable external memory to overwrite outdated knowledge. GRACE represents parameter-injection approaches by introducing new parameter matrices to encode knowledge with minimal disruption to original weights. We also include a lightweight prompt-based method that edits via input prompt without altering model parameters. Retrieval-augmented methods are excluded, as they rely on external sources rather than modifying internal representations, and thus fall outside the scope of intrinsic editing (Song et al. 2024; Shi et al. 2024; Chen et al. 2024b; Zheng et al. 2023).

Experiments

Experimental Settings

We conduct experiments using two state-of-the-art open-source MLLMs: `LLaVA-OneVision` and `Qwen2-VL`. Both models share the Qwen2-7B language backbone but differ significantly in their visual encoding architectures and multimodal alignment strategies. Building upon these models, we evaluate four representative knowledge editing paradigms: Prompt, LoRA, GRACE and WISE. To eliminate the confounding effects introduced by visual modality variations, all editing methods are applied exclusively to the language components, with the visual encoders kept frozen throughout the experiments. Additionally, to enhance the clinical relevance of our findings, we conduct experiments on a state-of-the-art medical-domain MLLM, `HuatuoGPT`, to validate the generalizability and robustness of editing methods under realistic medical scenarios.

To systematically assess the performance of knowledge editing methods under different levels of knowledge injection, we adopt two complementary evaluation settings. **Single Editing** focuses on the precise injection of an individual medical fact and evaluates the model's immediate response accuracy on associated queries immediately after the update, effectively reflecting the model's responsiveness to targeted knowledge modifications. **Lifelong Editing**, on the other hand, simulates continuous and incremental knowledge evolution in real-world applications by sequentially injecting 50, 250, 500, 750, and 1000 knowledge entries. This setup is used to examine the model's ability to retain newly added knowledge, mitigate cumulative interference, and robustly resist catastrophic forgetting during long-term editing.

Main Results

For Q1: How Do Current Knowledge-Editing Methods Perform Overall in Multimodal Medical Scenarios? Experimental results in multimodal medical scenarios (Table 2) reveal significant disparities in performance across knowledge-editing methods, driven by inherent trade-offs rooted in their design principles. Broadly, methods fall into two categories: external guidance approaches like Prompt, which injects knowledge via in-context learning without altering model parameters, and internal parametric methods such as WISE, LoRA, and GRACE. Prompt effectively leverages the model's language priors to achieve high Reliability and Generality, but their lack of structural constraint leads to poor Locality, causing interference with unrelated tasks. In contrast, parametric methods introduce targeted structural updates that ensure near-perfect Locality by isolating the edit's impact. However, their efficacy is inconsistent and generalization is limited. LoRA, which uses low-rank adapters, performs well on some models (e.g., 0.9033 on `HuaTuoGPT-7B`) but collapses on others (e.g., 0.1852 on `QWen2-VL`), exposing its dependency on model architecture. WISE relies on external memory, and its performance is tightly coupled with how well the memory module integrates into the model's reasoning pathways. GRACE, which injects new parameters, achieves strong reliability but suffers from poor generality (e.g., 0.4295), indicating brittle, over-localized edits that resemble memorization rather than knowledge integration. In essence, current methods face a structural dilemma: they either promote generalizable knowledge at the expense of control, or enforce localized precision while failing to support transferable reasoning.

For Q2: How Does Editing Effectiveness Differ Between the Understanding and Reasoning Tasks? A comparison between *Understanding* and *Reasoning* tasks (Table 2) reveals not a simple complexity gap, but fundamentally a mechanism-driven interaction between editing methods and task types. Methods like WISE and LoRA, which rely on internal parameter updates (e.g., memory injection or low-rank tuning), are effective for localized factual edits but struggle to generalize across temporally and multimodally complex reasoning inputs due to their scoped, non-generative nature. However, this is not universally true. GRACE, for instance, achieves near-perfect Reliability on Reasoning in `QWen2-VL` but underperforms on Understanding. Its injected parameters may act as shortcuts in the inference chain, enabling multi-step reasoning but failing to generalize due to low semantic abstraction. Similarly, Prompt performs better on Reasoning in `HuaTuoGPT-7B`, consistent with their design:prompt operates at the input level, and when the base model has strong inferential capabilities, context injection can more effectively propagate through reasoning paths than in factual recall. Ultimately, editing success hinges less on task difficulty and more on alignment between editing scope, cognitive demands, and model architecture. *Under-*

Methods	Task Type	LLaVA-Onevision				QWen2-VL				HuaTuoGPT-7B			
		Rel.	Gen.	T-Loc.	M-Loc.	Rel.	Gen.	T-Loc.	M-Loc.	Rel.	Gen.	T-Loc.	M-Loc.
Prompt	*Understanding*	0.9749*	0.9565*	0.7729	0.7986	0.8871*	0.5380*	0.7874	0.7367	0.0170	0.0236	0.8151	0.8099
	Reasoning	0.9413†	0.9234†	0.8055	0.8440	0.3897	0.4936†	0.7865	0.7522	0.0953	0.2693†	0.8076	0.8040
WISE	*Understanding*	0.5058	0.5122	1.0000*	1.0000*	0.5262	0.4593	1.0000*	1.0000*	0.8915*	0.8155*	1.0000*	1.0000*
	Reasoning	0.3997	0.3818	1.0000†	1.0000†	0.4599	0.4248	1.0000†	1.0000†	0.7550†	0.6956†	1.0000†	1.0000†
LoRA	*Understanding*	0.4546	0.4458	0.9987	0.9989	0.1852	0.1756	0.9990	0.9980	0.9033*	0.8871*	0.9748	0.9531
	Reasoning	0.3102	0.3209	0.9993	0.9981	0.0874	0.0788	0.9995	0.9951	0.8152†	0.7787†	0.9792†	0.9669
GRACE	*Understanding*	0.9173	0.4295	1.0000*	1.0000*	0.8827	0.1616	1.0000*	1.0000*	0.5417	0.0007	1.0000*	1.0000*
	Reasoning	0.2872	0.2973	1.0000†	1.0000†	0.9964†	0.0695	1.0000†	1.0000†	0.5294	0.0215	1.0000†	1.0000†

Table 2: Evaluation results of four editing methods under single edit. Asterisks (*) indicate the highest value per column in the *Understanding* scenario, and daggers (†) mark the highest in *Reasoning*.

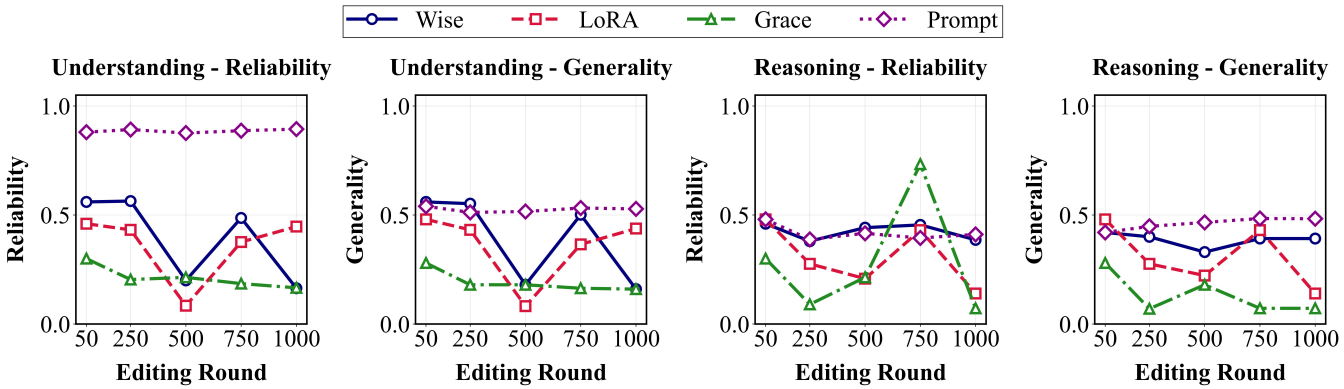


Figure 4: Lifelong editing results of Qwen2vl-7B using four methods: Prompt, WISE, GRACE, and LoRA. Left two subplots correspond to *understanding* tasks, and right two subplots correspond to *reasoning* tasks.

standing tasks require stable factual rewrites across variants, while Reasoning tasks demand edits that integrate effectively into dynamic inference mechanisms.

For Q3: How Does Edit Scale (Single Knowledge Editing vs. Lifelong Knowledge Editing) Affect Model Performance and Stability? Our investigation into continuous knowledge editing reveals that the primary challenge is not systematic performance degradation, but significant performance volatility and method-dependent stability as the number of edits increases, as shown in Figure 4. Regarding Locality, weight-modifying methods like WISE, GRACE, and LoRA consistently achieve near-perfect T-Locality and M-Locality scores (1.0), demonstrating a strong and consistent ability to prevent interference with unrelated knowledge, which is why these metrics were omitted from the visual plots for clarity. In sharp contrast, the prompt-based method shows persistently lower locality scores (around 0.75-0.80), indicating chronic and compounding side effects. However, a critical trade-off emerges in Reliability and Generality. The prompt-based method offers remarkable stability with constant, predictable performance. Conversely, WISE, GRACE, and LoRA suffer from extreme instability, with erratic performance featuring unpredictable peaks and troughs. This is starkly exemplified by GRACE's Reliability score, which spiked at 750 edits before collapsing. This highlights a fundamental dilemma: prompt-based methods provide stable but leaky edits with poor locality, whereas current weight-space methods offer well-contained but volatile edits whose success becomes chaotic and unreliable over time.

Efficiency Comparison

Time Consumption Analysis. As shown in Figure 6a, the variation in editing time directly reflects the structural complexity and computational overhead of different methods. The high latency observed in LoRA and WISE stems from their deep interventions in model parameters—LoRA introduces low-rank adapters, while WISE invokes external memory and re-encodes semantics. These intrusive operations significantly extend the computational path, especially in multimodal reasoning tasks. In contrast, Prompt performs localized and lightweight edits, relying solely on direct manipulation within the activation space, without reconstructing full computational graphs, thereby achieving faster response. GRACE, as a parameter-injection method, adopts a more modular structural design, resulting in intermediate latency. Overall, the more intrusive the editing mechanism, the harder it becomes to support low-latency updates required in time-sensitive scenarios such as clinical applications.

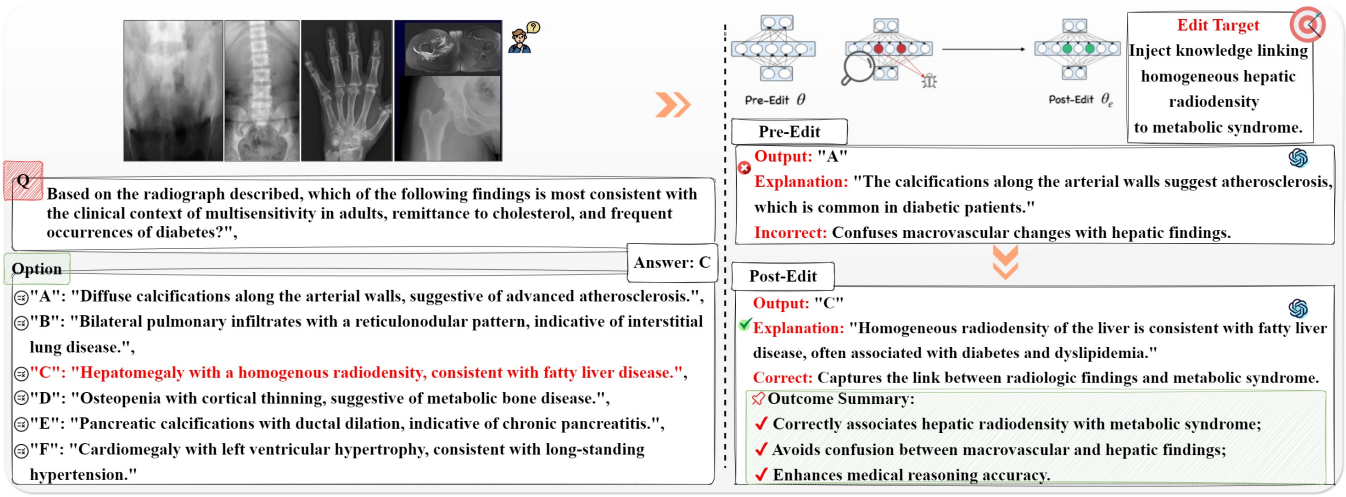
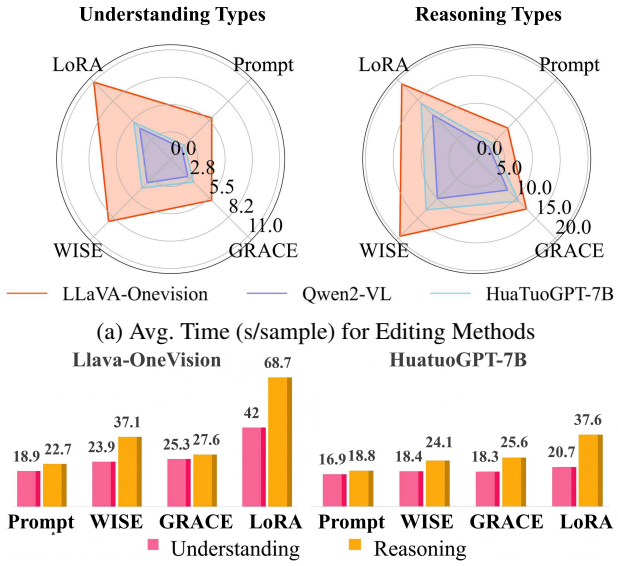


Figure 5: Case analysis of editing LLaVA-OneVision with WISE.



(a) Avg. Time (s/sample) for Editing Methods

(b) Peak GPU Mem. (GB/sample) for Editing Methods

Figure 6: Efficiency analysis of different editing methods.

Memory Consumption Analysis. Figure 6b shows that memory usage reflects differences in intermediate state storage and structural overhead. Although LoRA is parameter-efficient, it incurs the highest memory footprint due to the need to maintain adapter weights and gradient caches, especially when handling high-dimensional inputs. WISE and GRACE also show increases in memory consumption in multi-frame reasoning tasks, attributed to the inclusion of external modules or additional computational paths. In contrast, Prompt demonstrates the most stable memory profile, benefiting from its local editing strategy that avoids loading additional structures or states, thus exhibiting stronger resource adaptability. This makes Prompt more suitable for deployment in resource-constrained edge devices or real-world medical systems. In general, memory efficiency not only affects scalability but also reflects the resource-friendliness of the method's architectural design. Ultimately, these findings position the Prompt method as the most practical choice for scalable deployment.

Case Analysis

As illustrated in Figure 5, applying the WISE method to `LLaVA-OneVision` successfully injects the missing association between homogeneous hepatic radiodensity and metabolic syndrome. Initially, the model incorrectly selects a macrovascular diagnosis (Option A), failing to distinguish vascular calcifications. After the targeted edit, the model correctly selects Option C, demonstrating an understanding that hepatomegaly with homogeneous radiodensity is indicative of fatty liver disease—a manifestation commonly linked to various metabolic disorders. This shift reflects precise localization of the injected knowledge and improved multimodal reasoning, despite the edit being applied only to the language module. The overall outcome highlights the cross-modal coupling in vision-language models: linguistic edits can effectively propagate to correct grounded diagnostic reasoning without fine-tuning the vision backbone.

Conclusion and Future Outlook

In this paper, we present **MultiMedBench**, the first comprehensive benchmark for evaluating knowledge editing in multimodal medical scenarios. Systematic evaluations reveal that current editing methods exhibit significant limitations in accuracy, order sensitivity, and catastrophic forgetting, particularly in complex tasks requiring deep semantic understanding and robust clinical reasoning. While our benchmark provides a solid foundation, it is currently limited by a primary focus on question-answering tasks and lacks deeper insights into the interpretability of editing mechanisms. Future work aims to expand task complexity, enhance interpretability, and develop minimally invasive editing methods for sustainable long-term knowledge maintenance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62506166, U2441285, 62222605), the Natural Science Foundation of Jiangsu Province (No. BK20251365), the China Postdoctoral Science Foundation (No. 2025M774283). This research is also sponsored by the DiDi GAIA Collaborative Research Funds (No. CCF-DiDi GAIA202507) and CAAI-MindSpore Open Fund (CAAIXSJLJJ 2025 MindSpore 01), developed on OpenI Community.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; and et al. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Boger, E.; Alayrac, J.-B.; Menon, R.; and et al. 2023. Med-Flamingo: A Multimodal Medical Few-shot Learner. *arXiv preprint arXiv:2307.15195*.

Cai, Y.; Cao, D.; and et al. 2024. O-Edit: Orthogonal Subspace Editing for Language Model Sequential Editing. *arXiv preprint arXiv:2410.11469*.

Chen, J.; Gui, C.; Ouyang, R.; and et al. 2024a. HuatuoGPT-Vision: Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In *Proceedings of EMNLP*.

Chen, L.; Xie, Z.; Zhu, Y.; and et al. 2022a. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *arXiv preprint arXiv:2209.06794*.

Chen, Q.; Zhang, T.; He, X.; Li, D.; Wang, C.; Huang, L.; and Xue, H. 2024b. Lifelong Knowledge Editing for LLMs with Retrieval-Augmented Continuous Prompt Learning. *arXiv preprint arXiv:2405.03279*.

Chen, X.; Wang, C.; Xue, Y.; and et al. 2024c. Unified Hallucination Detection for Multimodal Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chen, X.; Wang, C.; Xue, Y.; and et al. 2024d. Unified Hallucination Detection for Multimodal Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chen, X.; Zhang, N.; Li, L.; and et al. 2022b. Hybrid Transformer with Multi-level Fusion for Multimodal Knowledge Graph Completion. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Chen, X.; Zhang, N.; Xie, X.; and et al. 2022c. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In *Proceedings of the ACM Web Conference (WWW)*.

Cheng, S.; Zhang, N.; Deng, S.; and Chen, H. 2024. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. *arXiv preprint arXiv:2501.18362*.

Dai, W.; Chen, P.; Lu, M.; Li, D.; Wei, H.; Cui, H.; Liang, P. P.; et al. 2025. Data Foundations for Large Scale Multimodal Clinical Foundation Models. *arXiv preprint arXiv:2503.07667*.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and et al. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*.

Dhingra, B.; Cole, J. R.; Eisenschlos, J. M.; and et al. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*.

Fang, J.; Jiang, H.; Wang, K.; and et al. 2024. AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models. *arXiv preprint arXiv:2410.02355*.

Feng, Y.; Zhan, L.; Lu, Z.; and et al. 2025. GeoEdit: Geometric Knowledge Editing for Large Language Models. *arXiv preprint arXiv:2502.19953*.

Han, T.; Kumar, A.; Agarwal, C.; and et al. 2024. MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models. *arXiv preprint arXiv:2403.03744*.

Hartvigsen, T.; Sankaranarayanan, S.; Palangi, H.; and et al. 2023. Aging with GRACE: Lifelong Model Editing with Discrete Key-Value Adaptors. In *Advances in Neural Information Processing Systems*.

He, e. a. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. *arXiv preprint arXiv:2401.13601*.

He, Y.; Zhang, Y.; Wu, Z.; and et al. 2023. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *arXiv preprint arXiv:2305.10415*.

Hu, E. J.; Shen, Y.; Wallis, P.; and et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, Y.; Li, T.; Lu, Q.; and et al. 2024. OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM. *arXiv preprint arXiv:2402.12049*.

Huang, C.; Guo, Y.; Wang, S.; and et al. 2024a. OmniMedVQA: Towards Generalist Multimodal Medical Visual Question Answering. *arXiv preprint arXiv:2402.09181*.

Huang, Z.; Wang, K.; Li, Y.; and et al. 2024b. ChatRad: Towards Radiologist-AI Collaborative Diagnosis via Multimodal LLMs. *arXiv preprint arXiv:2404.01481*.

Jiang, Y.; Wang, Y.; Wu, C.; and et al. 2024. Learning to Edit: Aligning LLMs with Knowledge Editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kalajdzievski, D. 2024. Scaling Laws for Forgetting When Fine-Tuning Large Language Models. *arXiv preprint arXiv:2401.05605*.

Kim, Y.; Jeong, H.; Chen, S.; and et al. 2025. Medical Hallucinations in Foundation Models and Their Impact on Healthcare. *arXiv preprint arXiv:2503.05777*.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; and et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*.

Li, C.; Wong, C.; Zhang, S.; and et al. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890*.

Li, Y.; Zhu, Y.; Yan, T.; Fan, S.; Wu, G.; and Xu, L. 2024. Knowledge Editing for Large Language Model with Knowledge Neuronal Ensemble. *arXiv preprint arXiv:2412.20637*.

Liu, Q.; Deng, S.; Zhang, N.; and Chen, H. 2024. MedFrameQA: A Multi-Image Medical VQA Benchmark for Clinical Reasoning. *arXiv preprint arXiv:2505.16964*.

Luo, Y.; Yang, Z.; Meng, F.; and et al. 2023. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. *arXiv preprint arXiv:2308.08747*.

Meng, K.; Bau, D.; Andonian, A.; and et al. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*.

Meng, K.; Sharma, A. S.; Andonian, A. J.; and et al. 2023. Mass-Editing Memory in a Transformer. In *International Conference on Learning Representations*.

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

Mitchell, E.; Lin, C.; Bosselut, A.; and et al. 2022b. Memory-Based Model Editing at Scale. In *International Conference on Machine Learning*.

Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2021. Fast Model Editing at Scale. *arXiv preprint arXiv:2110.11309*.

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 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 2056–2066.

Song, M.; Wang, Z.; He, K.; Dong, G.; Mou, Y.; Zhao, J.; and Xu, W. 2024. Knowledge Editing on Black-Box Large Language Models. *arXiv preprint arXiv:2402.08631*.

Wang, P.; Li, Z.; Zhang, N.; and et al. 2024a. WISE: Rethinking the Knowledge Memory for Lifelong Model Editing of Large Language Models. *arXiv preprint arXiv:2405.14768*.

Wang, X.; Zhang, N. X.; He, H.; and et al. 2024b. Safety Challenges of AI in Medicine in the Era of Large Language Models. *arXiv preprint arXiv:2409.18968*.

Wu, W.; Xu, X.; Gao, C.; Diao, X.; Li, S.; Salas, L. A.; and Gui, J. 2025. Assessing and Mitigating Medical Knowledge Drift and Conflicts in Large Language Models. *arXiv preprint arXiv:2505.07968*.

Wu, X.; Bu, Y.; Cai, Y.; and Wang, T. 2024. Updating Large Language Models' Memories with Time Constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 13693–13702.

Xu, D.; Zhang, Z.; Zhu, Z.; and et al. 2024a. Editing Factual Knowledge and Explanatory Ability of Medical Large Language Models. *arXiv preprint arXiv:2402.18099*.

Xu, Z.; Jiang, X.; Liu, Y.; Liu, L.; Tang, B.; Xu, H.; and Wang, X. 2024b. MedKEBench: A Comprehensive Benchmark for Knowledge Editing in Medical Large Language Models. *arXiv preprint arXiv:2506.03490*.

Yan, Q.; He, X.; Yue, X.; and Wang, X. E. 2024. Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA. *arXiv preprint arXiv:2405.20421*.

Yang, Z.; Li, L.; Lin, K.; and et al. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*.

Yao, Y.; Fang, J.; Gu, J.-C.; and et al. 2025. CaKE: Circuit-aware Editing Enables Generalizable Knowledge Learners. *arXiv preprint arXiv:2503.16356*.

Yao, Y.; Wang, P.; Tian, B.; and et al. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Youssef, P.; Zhao, Z.; Seifert, C.; and Schlötterer, J. 2025. Tracing and Reversing Rank-One Model Edits. *arXiv preprint arXiv:2505.20819*.

Yu, L.; Chen, Q.; Zhou, J.; and et al. 2024. MELO: Enhancing Model Editing with Neuron-Indexed Dynamic LoRA. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhang, F.; Li, Q.; Sun, H.; and et al. 2023a. XrayGPT: Chest Radiograph Diagnosis with Language-Image Pre-training and Adaptation. *arXiv preprint arXiv:2307.01850*.

Zhang, N.; Yao, Y.; Tian, B.; and et al. 2024. A Comprehensive Study of Knowledge Editing for Large Language Models. *arXiv preprint arXiv:2401.01286*.

Zhang, Q.; Chen, M.; Bukharin, A.; He, P.; Cheng, Y.; Chen, W.; and Zhao, T. 2023b. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *International Conference on Learning Representations (ICLR)*. Available at https://arxiv.org/abs/2303.12520.

Zhang, Q.; Chen, M.; Bukharin, A.; He, P.; Cheng, Y.; Chen, W.; and Zhao, T. 2023c. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *The Eleventh International Conference on Learning Representations*.

Zheng, C.; Li, L.; Dong, Q.; and et al. 2023. Can We Edit Factual Knowledge by In-Context Learning? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.