This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Can Knowledge be Transferred from Unimodal to Multimodal? Investigating the Transitivity of Multimodal Knowledge Editing

Lingyong Fang[1,2*], Xinzhong Wang[1,2*], Depeng Wang[2], Zongru Wu[1], Ya Guo[2],
Huijia Zhu[2†], Zhuosheng Zhang[1], Gongshen Liu[1,3†]
[1]Shanghai Jiao Tong University, China [2]Ant Group, China
[3]Inner Mongolia Research Institute of SJTU, China
{fangly, 2046449167, wuzongru, zhangzs, lgshen}@sjtu.edu.cn
{wdp432379, guoya.gy, huijia.zhj}@antgroup.com

Abstract

Multimodal Large Language Models (MLLMs) contain a substantial amount of factual knowledge, which may become outdated or inaccurate over time. Consequently, various knowledge editing techniques have been proposed to update the knowledge encoded within these models. Previous approaches maintain modality consistency during both the editing and testing phases. However, in practical applications, it is desirable for knowledge to be transferable across different modalities, which can enhance the robustness of knowledge editing and potentially allow for cost-effective editing of multimodal knowledge using textual information. To address this, we introduce the concept of Transitivity of Multimodal Knowledge Editing (TMKE) and design corresponding evaluation criteria. Subsequently, we construct a corresponding TMKE Benchmark through an automated pipeline. We evaluate three MLLMs and five knowledge editing methods, uncovering limitations in the current models and methods concerning transitivity. Additionally, we analyze the intrinsic representations of the model during the editing process based on Knowledge Neurons to interpret the experimental phenomena.

1. Introduction

Multimodal Large Language Models (MLLMs) [3, 10, 12, 22] have achieved remarkable performance across a variety of downstream multimodal tasks. However, the knowledge stored within these models may become outdated or contain errors, which can impact downstream applications in both unimodal and multimodal settings. Consequently, knowledge editing within LLMs has gained increasing attention. The core objective is to modify such information within the models effectively without requiring retraining from scratch, while ensuring that other knowledge remains unaffected.

Previous studies have proposed several knowledge editing methods designed to inject textual knowledge into Large Language Models (LLMs), such as MEND [15] and ROME [14]. Inspired by these techniques within LLMs, some papers have introduced these methodologies to MLLMs. The common thread among these studies is that they all directly inject knowledge at the multimodal level; in other words, the data used for editing is inherently multimodal, including both images and text. However, multimodal data is far less abundant than unimodal data, making knowledge editing with multimodal data resource-intensive. Consequently, the question of whether unimodal knowledge can be transferred to multimodal scenarios has become significant. This implies that there is potential to leverage unimodal data for more cost-effective knowledge editing in multimodal contexts.

To evaluate the effectiveness of knowledge editing methods [14–16, 20], previous works [4, 7, 11] have proposed various unimodal and multimodal knowledge editing benchmarks. However, these efforts are disjointed, implying that the performance of unimodal editing models in the multimodal context remains unknown. With the prevalence of MLLMs, an increasing number of models deployed in real-world applications are multimodal rather than unimodal. This suggests that direct knowledge editing on unimodal models is less common. Consequently, investigating the correlation of knowledge editing performance across different modalities has become increasingly important.

To address the aforementioned challenges, we investigated Transitivity of Multimodal Knowledge Editing (TMKE). We propose that knowledge editing should transcend the single modality being edited and transmit across modalities. Figure 1 illustrates an example of TMKE. To in-

*Equal contribution. † Corresponding author.

corporate transitivity into knowledge editing, we defined the problem and proposed corresponding evaluation criteria to test which facts, beyond the modality being edited, should be modified or retained post-editing. Building on this, we created the TMKE Benchmark (TMKEB), a new comprehensive benchmark for evaluating MLLMs knowledge editing capabilities. Initially, we leveraged the MQuAKE dataset [21], extending it to a multimodal scenario through an automated pipeline. Subsequently, we continued to expand the relevant questions and images in accordance with the evaluation criteria.

We utilized TMKEB to evaluate three popular MLLMs and five commonly-used knowledge editing methods. Our findings indicate that: (1) current knowledge editing techniques generally exhibit high success rates in unimodal editing, yet there is a notable decline for some methods when assessed through transitivity metrics; (2) the evaluation of the impact on unrelated knowledge reveals that locality, previously widely used, is insufficient for accurate measurement. Our newly proposed stability metric uncovers that existing knowledge editing methods significantly affect unrelated knowledge; (3) all current knowledge editing methods demonstrate clear limitations when evaluated through transitivity metrics; (4) the FT and ROME methods present contrasting characteristics regarding transitivity, with FT showing higher editing success rates while ROME exhibits less impact on unrelated knowledge. To further analyze these experimental phenomena, particularly why existing knowledge editing methods perform poorly in transitivity metrics, and the seesaw characteristics presented in the reliability and stability metrics, Knowledge Neurons (KN) [5] were employed to analyze the models. We formulated and addressed five research questions to delve deeper into these issues.

In summary, our main contributions are as follows: (1) We propose the Transitivity of Multimodal Knowledge Editing (TMKE) and design corresponding evaluation criteria. (2) We introduce the TMKE Benchmark and conduct a comprehensive evaluation of existing knowledge editing methods. Experimental results indicate that current methods still face challenges concerning transitivity. (3) Knowledge Neurons (KN) are utilized to analyze the models during the knowledge editing process, providing explanations for the observed experimental phenomena.

2. Related Work

2.1. Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) are typically constructed by integrating a visual encoder with a language model, connected through an intermediary module. Alayrac et al. [1] introduced an innovative method employing a query-based cross-attention mechanism, resulting in a robust vision-language interaction module. BLIP-2 [10] replaced cross-attention with Q-Former, a streamlined Transformer-based architecture. Enhancements in models like MiniGPT-4 [22] stem from leveraging instruction-tuning datasets compiled from diverse public sources. LLaVA [12] and Otter [9] developed an instruction data system to improve comprehension capabilities. Additionally, Bai et al. [3] introduced a three-stage training framework to better align multimodal representations, offering further advancement over previous methods.

2.2. Knowledge Editing

For unimodal knowledge editing, various benchmarks have been introduced to address this task, such as COUNTERFACT [14] for counterfactual knowledge, MQuAKE [21] for multi-hop reasoning, AToKE [19] for preserving prior knowledge, and WIKIUPDATE [18] for handling unstructured knowledge.

For multimodal knowledge editing, MMEdit [4] was introduced, building upon Visual QA [2] and Image Captioning [6]. MIKE [11] targets the precise editing of multimodal entity knowledge. Huang et al. [7] proposed VLKEB, which leverages multimodal knowledge graphs (Liu et al., 2019) to assess vision knowledge editing.

However, previous work has been limited to a fixed editing scenario, without exploring transitive knowledge editing which conducts text-based knowledge editing but testing with multimodal knowledge. Additionally, previous work lacked discussions on similar issues, which is the focus of our research.

3. Transitivity of Multimodal Knowledge Editing

3.1. Problem Setting

We consider the editing of factual knowledge, where facts are expressed as triplets (s, r, o), consisting of a subject s (such as the book title "The French Lieutenant's Woman"), a relationship r (such as "author"), and an object o (such as the person's name "William Gibson"). Depending on the modality, the subject can be categorized into a textual modality subject s_T and an image modality subject s_I. Evaluation of knowledge editing involves two phases: editing and testing. In the editing phase, certain knowledge is injected into the model, followed by the testing phase where the accuracy of other related or unrelated knowledge is evaluated.

By classifying existing knowledge editing evaluation methods based on the modality used in both the editing phase and the testing phase, we identify three categories of knowledge editing problems, with specific examples referred to in Figure 1.

Unimodal Knowledge Editing. Both the editing phase

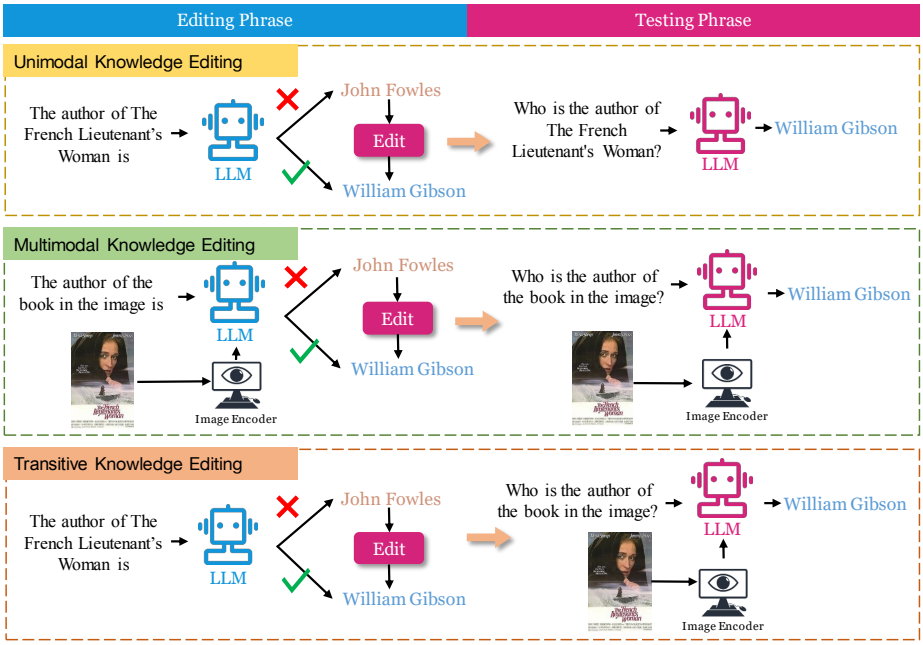

Figure 1. An illustrative example of TMKE.

and the testing phase are conducted in the pure text modality, i.e., the knowledge in the editing phase is (s_T, r, o) and in the testing phase is (s_T, r, o).

Multimodal Knowledge Editing. Both the editing phase and the testing phase are conducted in the image modality, i.e., the knowledge in the editing phase is (s_I, r, o) and in the testing phase is (s_I, r, o).

Transitive Knowledge Editing. The editing phase and the testing phase are conducted in different modalities, i.e., the knowledge in the editing phase is (s_T, r, o) and in the testing phase is (s_I, r, o). Conversely, editing can also be performed on images while testing is done on pure text. However, this scenario is impractical due to the significantly smaller amount of multimodal data compared to pure text data. In real-world scenarios, the direction of knowledge transfer should be from unimodal to multimodal; thus, this paper does not consider such scenarios.

3.2. Evaluation Criteria

The challenges in knowledge editing lie in:(1) ensuring that knowledge associated with the edited information is also correctly addressed; (2)maintaining the integrity of knowledge unrelated to the edits.

Consequently, during the testing phase, it is necessary not only to evaluate the (s, r, o) knowledge that was edited but also to assess its impact on other knowledge. The consideration of transitivity further complicates the evaluation. We have designed metrics for knowledge editing under the context of transitivity. The process of constructing datasets related to these metrics is detailed in Section 4, with examples provided in Fig 2. Let us assume the knowledge used during the editing phase is (s_T, r, o). Below, we describe the knowledge tested by each evaluation criterion and the significance of each criterion.

Reliability. The knowledge tested is (s_I, r, o). This is the most fundamental test of the knowledge being edited, ensuring that the edited knowledge itself can be accurately answered.

Generality. The knowledge tested is still (s_I, r, o). To prevent the model from simply memorizing a simple mapping between questions and answers, we need to vary the phrasing of the question and continue testing whether the edited knowledge can be correctly answered.

Portability. The knowledge tested is (s_I, r, o'). Our aim is for the model to truly understand the edited knowledge, so for multi-hop knowledge (s_I, r, o') related to (s_I, r, o), we expect the model to provide correct answers.

Locality. The knowledge tested is (s_I^*, r^*, o^*). After editing, the performance on knowledge entirely unrelated to the edits should remain unaffected; thus, we need to test knowledge completely unrelated to the edited knowledge.

Stability. The knowledge tested is (s_I^*, r, o^*). While Locality can measure the impact of knowledge editing on unrelated knowledge to some extent, the design of questions that are completely unrelated to the edited knowledge may make the task too easy. Therefore, we should also test knowledge with similar question forms but entirely different answers. This can help assess whether the model is answering questions correctly based on shortcuts derived from the form of the questions.

4. TMKE Benchmark

We conducted knowledge editing experiments across three distinct scenarios and introduced new evaluation metrics. To provide a more comprehensive assessment of the effectiveness of various methods and address the limitations of prior benchmarks, we present the TMKE Benchmark. The entire process of constructing our dataset can be referenced in Figure 2.

4.1. Data Generation Pipeline

4.1.1. Data Sources and Selection

The TMKE Benchmark was constructed by expanding from purely textual information to multimodal data, using MQuAKE [21] as the foundation. Zhong et al. [21] developed MQuAKE as a textual multi-hop knowledge editing dataset, consisting of knowledge triples $t = (s, r, o)$ and associated multi-hop questions.

4.1.2. Data Construction

Data Modality Expansion. Since MQuAKE is composed solely of textual information, we expanded it by incorporating visual data. Using the subject s from the knowledge triples $t = (s, r, o)$, we input its tokens as search keywords into the Google[1] search engine. Images from the returned URLs were manually filtered, and for each subject s, we selected the five most relevant images to represent its multimodal context.

Reliability Dataset Construction. MQuAKE dataset provides the foundation for unimodal knowledge editing tasks. For transitive and multimodal knowledge editing, we randomly select two images from the set of five. One image is used during the editing process, while the other, unseen by the model, assesses post-editing robustness.

To adapt textual descriptions for multimodal scenarios, we reframe questions accordingly. For instance, for the question, "Who is the author of The French Lieutenant's Woman?", we first use GPT-4o [17] to identify entity types, recognizing "The French Lieutenant's Woman" as a "book". We then modify the question to: "Who is the author of the book in the image?" The specific prompt for GPT-4o is provided in Appendix.

Generality Dataset Construction. Knowledge editing is performed using a question-answer format, aiming for the model to correctly answer rephrased versions of the original questions. For unimodal editing, GPT-4o rephrases the questions while preserving their meaning (see Appendix). For transitive and multimodal tasks, we ensure consistency with the images in the Reliability dataset and adjust textual descriptions to align with the multimodal context.

[1] https://www.googleapis.com/customsearch/v1

Portability Dataset Construction. Given that MQuAKE is a multi-hop dataset, we design our textual editing data directly based on its structure. For multimodal tasks, we ensure consistency by using the same images as in the Reliability dataset and adapting textual descriptions accordingly.

Locality Dataset Construction. To ensure that the model's unrelated knowledge remains intact before and after the editing process, we follow the settings from MMEdit [4]. We utilize the NQ [8] dataset for textual editing and the OK-VQA [13] dataset for multimodal editing. KL scatter plots are generated to compare the model's pre-edit and post-edit outputs, further constraining the editing process.

Stability Dataset Construction. We modify the model's knowledge triples from (s, r, o) to (s, r, o') but expect consistent model responses for questions involving the same relation with different instances (s', r, o''). For textual editing, GPT-4o generates a new subject s' of the same type and a corresponding answer o'' (filtered for accuracy). The prompt for GPT-4o is in Appendix. For transitive and multimodal tasks, we retrieve images of s' via Google search, following the same procedure as in the Reliability dataset, and adjust textual descriptions for the multimodal context.

4.2. Data Statistics

Eventually, we created the TMKE Benchmark, which comprises a total of 3,676 examples, including 2,352 training examples and 1,324 test examples. TMKE Benchmark covers 101 types of entities, with two images corresponding to each subject s, demonstrating the diversity of our dataset.

5. Experiments

We conducted extensive experiments on the TMKE Benchmark, covering various editing scenarios, methods, and multimodal models. Our analysis evaluated the effectiveness of each editing scenario and compared the performance of different methods. Specific experimental parameters and settings are detailed in Appendix.

5.1. Experiments Setup

We experimented with five knowledge editing methods: **FT**, **IKE** [20], **MEND** [15], **SERAC** [16], and **ROME** [14]. Additionally, we tested three MLLMs: **BLIP2** [10], **MiniGPT-4** [22], and **LLaVA-1.5** [12]. Detailed descriptions of these models and methods are provided in Appendix. The **Metrics** used in our experiments are also referenced in Appendix.

5.2. Results and Analysis

We conducted a comparative analysis of the various editing methods across different scenarios using the **TMKE** Benchmark. Detailed results for **Unimodal Knowledge Editing**,

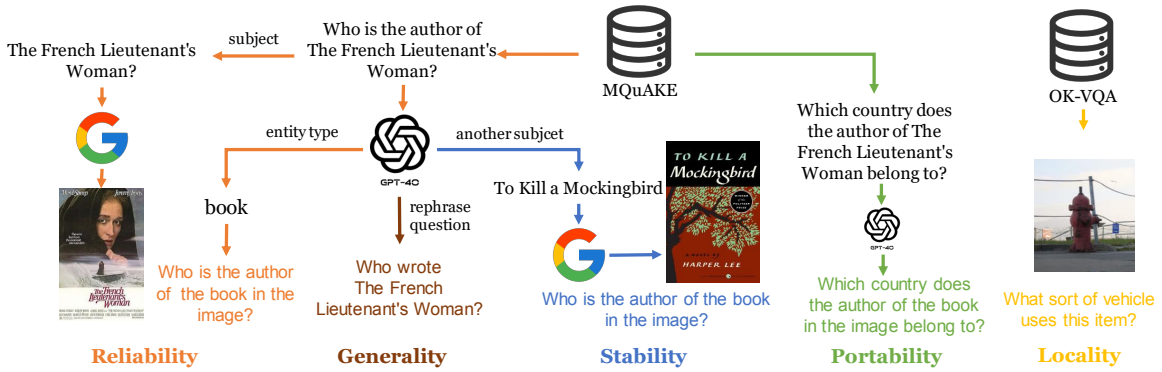

Figure 2. Overview of the TMKE Dataset Construction Process.

Transitive Knowledge Editing, and **Multimodal Knowledge Editing** are presented in Tables 1, 2, and 3. Due to space constraints, and given that the experimental results on BLIP2-OPT and MiniGPT-4 are largely consistent with those of the LLaVA-1.5 model, we have only presented the results of the LLaVA-1.5 model here. The complete experimental results can be found in the Appendix.

Reliability. In unimodal knowledge editing, all methods achieved a Reliability score above 90%, indicating strong performance. Parameter-update methods like FT, MEND, and ROME effectively incorporated new knowledge during edits. Memory-based methods, such as SERAC, identified and appended the most relevant information to the test inputs, while IKE introduced the edit instance as a "New Fact" prompt before each test, also resulting in high Reliability.

In transitive knowledge editing, FT and IKE maintained good performance, but MEND, SERAC, and ROME showed significant declines. MEND and SERAC, which require additional training, struggled with visual information, as their extra parameters trained on textual data were inadequate for this task. ROME's performance drop will be discussed further in Section 6. In multimodal knowledge editing, where the training and test data modalities aligned, all methods demonstrated better Reliability compared to transitive knowledge editing.

Generality. Owing to the inherent generalization abilities of Large Language Models (LLMs), performance on rephrased questions remained similar to Reliability. In transitive knowledge editing, where the images were unseen during the editing process, Generality remained stable when testing with different images. However, in multimodal knowledge editing, **Generality_U** declined compared to **Generality_S**, indicating that the model tends to associate specific images with certain outcomes, leading to reduced accuracy when different images are used.

Portability. Across all scenarios, Portability did not exceed 50%, suggesting that current editing methods struggle to transfer edited knowledge effectively in reasoning tasks. In unimodal knowledge editing, memory-based methods like SERAC and IKE outperformed parameter-update methods (FT, MEND, ROME), implying that supplying new knowledge directly is more effective than encoding it into parameters for solving multi-hop problems. Portability declined in transitive and multimodal knowledge editing, highlighting the limitations of current methods in handling multimodal multi-hop tasks.

Locality. Locality performance varied among methods. In unimodal knowledge editing, MEND, SERAC, and ROME performed well, while FT, which fine-tunes the final layer to fit the edited answer, showed reduced accuracy on unrelated questions. IKE, which explicitly provides new knowledge as a prompt, performed worst on Locality, as even unrelated questions caused the model to focus on the edited knowledge.

In transitive and multimodal editing, Locality_T remained similar to unimodal knowledge editing. However, for memory-based methods like IKE and SERAC, Locality_I was significantly lower than Locality_T. This might be due to SERAC's classifier misidentifying out-of-scope examples, as its counterfactual model is tailored for in-scope knowledge recovery. IKE's low Locality_I is likely caused by the appended edited facts misleading the model when testing on new information.

Stability. While both Locality and Stability measure the model's resistance to unrelated knowledge, Stability values were consistently lower. When faced with unrelated questions, the model can often provide the correct answer. However, when only the entity changes in the question, the model frequently confuses the edited entity with the new one, leading to incorrect answers.

Model	Method	Reliability	Generality	Portability	Locality	Stability
LLaVA-1.5 (7B)	FT	100.0	100.0	20.3	94.5	32.3
	IKE	100.0	99.8	48.7	60.0	56.2
	MEND	98.6	98.4	25.3	99.5	58.3
	SERAC	90.0	84.5	25.5	100.0	21.7
	ROME	97.8	95.6	34.1	98.9	62.1

Table 1. The unimodal knowledge editing results of various editing methods applied to LLaVA-1.5. The best results are indicated in a dark color, and the second-best results are indicated in a light color.

Model	Method	$\text{Reliability}_\text{S}$	$\text{Reliability}_\text{U}$	$\text{Generality}_\text{S}$	$\text{Generality}_\text{U}$	Portability	Locality_T	Locality_I	Stability
LLaVA-1.5 (7B)	FT	99.3	99.2	99.2	99.1	21.3	84.5	38.6	30.7
	IKE	100.0	100.0	95.1	94.9	24.9	61.8	1.0	20.2
	MEND	67.9	68.6	65.7	66.2	24.5	99.5	89.3	46.1
	SERAC	26.1	26.2	26.0	25.9	13.7	100.0	10.7	14.4
	ROME	30.6	30.7	30.9	31.1	24.1	98.7	88.5	57.9

Table 2. The transitive knowledge editing results of various editing methods applied to LLaVA-1.5. The best results are indicated in a dark color, and the second-best results are indicated in a light color.

Current editing methods struggle to isolate knowledge edits effectively. Non-essential elements like question phrasing still influence results. In multimodal knowledge editing, Stability was notably lower than in transitive knowledge editing, except for IKE. IKE, which constructs prompts using text in all scenarios, exhibited more consistent performance. For FT, MEND, and SERAC, the main distinction between scenarios was whether the image was visible during editing. In multimodal knowledge editing, the model often linked specific inputs to specific answers, increasing the likelihood of repeating the edited answer for similar questions—consistent with the Generality observations.

Comparative Analysis. In transitive knowledge editing, the model is edited using textual information but tested with images. In multimodal knowledge editing, both editing and testing use multimodal data and this consistency leads to higher Reliability. However, differences across images, as reflected in the Generality metric, and lower Stability in multimodal knowledge editing indicate that the model tends to form direct associations between specific inputs and answers. When faced with similar questions or images, the model often provides the edited answer rather than the correct one. The causes of this behavior will be explored further in Section 6.

6. Exploring Knowledge Neurons in Multimodal Knowledge Editing

We define a knowledge triple as $t = (s, r, o)$, where the multimodal model infers this knowledge via the process $I \rightarrow S \rightarrow O$. Here, I denotes the multimodal input provided to the model, S represents the subject, and O is the corresponding knowledge object. The process $I \rightarrow S$ aligns the multimodal input I with the subject S using components like the visual encoder, while $S \rightarrow O$ maps the subject to the relevant knowledge stored in the model. In transitive knowledge editing, the process mirrors unimodal editing by altering $S \rightarrow O$ to $S \rightarrow O'$. In contrast, direct multimodal knowledge editing modifies the entire sequence from $I \rightarrow O$ to $I \rightarrow O'$.

The findings from Section 5 revealed several significant patterns. For example, while knowledge editing performs well in unimodal scenarios, FT retains its performance in transitive knowledge editing, whereas ROME experiences a marked decline. Furthermore, Stability tends to be higher in transitive knowledge editing than in multimodal knowledge editing.

To investigate these observations further, we leveraged the experimental results and applied the Knowledge Neurons framework [5] to analyze the model's internal behavior during reasoning. This section addresses the following research questions:

RQ1: Is the activation of knowledge neurons correlated with Reliability?

RQ2: Why do knowledge editing methods perform poorly in terms of Stability?

RQ3: In transitive knowledge editing, why does FT exhibit higher Reliability but lower Stability compared to ROME?

RQ4: Why is Stability higher in transitive knowledge editing than in multimodal knowledge editing?

RQ5: Does knowledge editing create shortcuts in VQA tasks?

Model	Method	$\text{Reliability}_\text{S}$	$\text{Reliability}_\text{U}$	$\text{Generality}_\text{S}$	$\text{Generality}_\text{U}$	Portability	Locality_T	Locality_I	Stability
LLaVA-1.5 (7B)	FT	100.0	100.0	100.0	97.7	21.1	80.3	43.2	6.5
	IKE	100.0	100.0	99.4	99.8	27.6	62.3	1.0	22.1
	MEND	97.8	98.3	95.4	90.2	29.8	92.1	85.9	14.4
	SERAC	50.7	51.3	49.8	34.9	17.7	100.0	5.1	0.7

Table 3. The multimodal knowledge editing results of various editing methods applied to LLaVA-1.5. The best results are indicated in a dark color, and the second-best results are indicated in a light color.

6.1. Knowledge Neurons

To measure disparities, we use Knowledge Neurons (KN), as introduced by Dai et al. [5], to identify key neurons activated in LLMs during their responses to various queries. Specifically, each neuron $w_i^{(l)}$ in the feedforward network (FFN) layer is gradually adjusted from 0 to its original value $\overline{w}_i^{(l)}$, while the gradient is integrated. The continuous integrals are approximated using the Riemann method, leading to the following expression:

$$\tilde{\text{Attr}}(w_i^{(l)}) = \frac{\overline{w}_i^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial P\left(\frac{k}{m}\overline{w}_i^{(l)}\right)}{\partial w_i^{(l)}}, \tag{1}$$

where $P(w_i^{(l)}) = p(y|x, w_i^{(l)} = \hat{w}_i^{(l)})$ represents the probability that the LLM generates the correct answer when neuron $w_i^{(l)}$ is set to $\hat{w}_i^{(l)}$, and m is the number of approximation steps. Neurons with attribution values above a threshold v are considered critical for the LLM's decision-making:

$$\mathbf{N} = \left\{ w_i^{(l)} \mid \text{Attr}(w_i^{(l)}) > v \right\}. \tag{2}$$

In this study, we set $m = 10$ and the attribution threshold $v = 0.15$.

We decompose the multimodal process $I \to O$ into two sub-processes: $I \to S \to O$. We analyze the neural activations triggered by the model when processing the full query $I \to O$, denoted as N_{IO}, as well as the neural activations for the sub-processes $I \to S$ and $S \to O$, denoted as N_{IS} and N_{SO}, respectively. For example, for the query $I \to O$, "Who is the author of the book in the image?" with an image of "The French Lieutenant's Woman", the subproblem $I \to S$ would be "What is the book in the image?" using the same image, and the subproblem $S \to O$ would be "Who is the author of The French Lieutenant's Woman?" asked in a text-based format.

We compare these three sets of neural activations to derive insights aligned with specific research questions. The details of these comparisons are outlined in the relevant sections.

6.2. RQ1: Is the activation of knowledge neurons correlated with Reliability?

Experiment Setup. Transitive knowledge editing focuses on the $S \to O$ process, while our evaluation tests the $I \to O$ process. A simple hypothesis is that if the $S \to O$ process significantly overlaps with the $I \to O$ process, then modifying $S \to O$ to $S \to O'$ will likely result in changing $I \to O$ to $I \to O'$, thereby improving the Reliability of transitive knowledge editing. We define this overlap as $M1$:

$$M1 = \frac{N_{SO} \cap N_{IO}}{N_{IO}}. \tag{3}$$

We calculate $M1$ across the dataset and divide the data into two subsets: D_1 with $M1 \geq 0.15$ and D_2 with $M1 \leq 0.05$. We then evaluate ROME's accuracy on these subsets.

Dataset	Method	M1	M3	Reliability
D_1	FT	0.2283	0.0308	0.9977
	ROME	0.1807	0.0642	0.7574
D_2	FT	0.2075	0.0218	0.9931
	ROME	0.0446	0.0277	0.0152

Table 4. The results of ROME and FT on datasets D_1 and D_2.

Dataset	Method	M2	Reliability	Stability
D_3	ROME	0.6530	0.7877	0.1027
D_4	ROME	0.4661	0.3508	0.6653

Table 5. The results of ROME on datasets D_3 and D_4.

Scene	Method	M2	Reliability	Stability
Transitive Knowledge Editing	FT	0.6439	0.9989	0.1130
Multimodal Knowledge Editing	FT	0.8431	1.0000	0.0157

Table 6. The results in transitive and multimodal knowledge editing.

Results. The results, shown in Table 4, indicate that a higher $M1$ correlates with better Reliability. This suggests that the activation of knowledge neurons is indeed related to Reliability.

6.3. RQ2: Why do knowledge editing methods perform poorly in terms of Stability?

Experiment Setup. Tables 2 and 3 show that a high Reliability score for a knowledge editing method does not al-

ways correspond to high Stability. To investigate, we split the data into two subsets: D_3, where the Stability questions are answered with the edited response, and D_4, where the Stability questions are answered correctly. We measure neural activations while answering Stability questions $I' \rightarrow S' \rightarrow O'$ and their subquestions $I' \rightarrow S'$ and $S' \rightarrow O'$, and compare them to activations during the edited query $I \rightarrow S \rightarrow O$ and its subcomponents. We define $M2$ as:

$$M2 = \frac{N_{I'O'} \cap N_{IO}}{N_{IO}}. \tag{4}$$

Results. As shown in Table 5, higher $M2$ correlates with higher Reliability but lower Stability. When $M2$ is high, neuron activation during Stability questions overlaps significantly with edited question responses, indicating that the model memorizes the edited answers, resulting in lower Stability. This suggests that Reliability alone is not a complete measure of editing effectiveness, as Stability provides a better gauge of the model's ability to handle variations in similar queries.

6.4. RQ3: In transitive knowledge editing, why does FT exhibit higher Reliability but lower Stability compared to ROME?

Experiment Setup. Table 2 shows that FT has higher Reliability but lower Stability than ROME in transitive knowledge editing. To investigate, we apply FT editing to the D_1 and D_2 datasets from RQ1, measuring both Reliability and Stability alongside $M1$ values for the edited model.

Results. Table 5 reveals that ROME exhibits high $M1$ in one dataset, leading to high Reliability. FT shows high $M1$ across both datasets, indicating that in transitive knowledge editing, $S \rightarrow O$ accounts for a larger portion of $I \rightarrow S \rightarrow O$, thus increasing Reliability for FT. However, this strong alignment also makes FT more likely to return the edited answer in Stability tests, reducing Stability. This reinforces the link between neuron activation and both Reliability and Stability.

6.5. RQ4: Why is Stability higher in transitive knowledge editing than in multimodal knowledge editing?

Experiment Setup. Comparing Tables 2 and 3, we observe that Stability in multimodal knowledge editing is generally lower than in transitive editing, except for IKE. To explore this, we use the FT method and the D_3 dataset from RQ2, where the model provided edited answers. We measure Reliability, Stability, and $M2$ after performing knowledge editing in both transitive and multimodal scenarios.

Results. Table 6 shows that both transitive and multimodal knowledge editing achieve high Reliability with FT. However, $M2$ is higher for multimodal editing, indicating that it leads the model to memorize the edited question-answer pairs, thus reducing Stability when similar questions are encountered. Multimodal editing increases Reliability but sacrifices the model's flexibility in addressing variations, while transitive editing achieves a more balanced performance.

6.6. RQ5: Does knowledge editing create shortcuts in VQA tasks?

Experiment Setup. In RQ1, we defined $M1$ as the proportion of $S \rightarrow O$ in the $I \rightarrow S \rightarrow O$ reasoning process. Here, we examine how much the image processing step $I \rightarrow S$ contributes to multimodal knowledge reasoning. We define $M3$ as:

$$M3 = \frac{N_{IS} \cap N_{IO}}{N_{IS}}. \tag{5}$$

We calculate $M3$ for models edited by FT and ROME on D_1 and D_2 datasets from RQ1.

Results. As shown in Table 4, FT, which has higher Reliability, shows a lower $M3$ value than ROME, indicating that FT relies less on image information during multimodal reasoning. In transitive knowledge editing, the model edits knowledge based on text but is tested using multimodal inputs. FT establishes a direct link between text input and the answer, bypassing image information, which boosts Reliability. However, this creates shortcuts, reducing Stability, as FT relies more on text than on integrating multimodal data.

7. Conclusion

In the field of knowledge editing, we have introduced the concept of Transitivity of Multimodal Knowledge Editing (TMKE), positing that the modification of specific facts implies further updates to related knowledge across other modalities. We have established evaluation criteria for TMKE and created the TMKE Benchmark, designed to evaluate the capability of knowledge editing methods in terms of transitivity. Our evaluation encompasses three Multimodal Large Language Models (MLLMs) and five knowledge editing (KE) methods, with results indicating limitations in transitivity among these approaches. This suggests that future developments in KE methodologies should more carefully consider the principle of transitivity. Finally, we provide a deeper insight into the experimental phenomena through the lens of Knowledge Neurons, revealing the underlying reasons for the current poor performance of KE methods with respect to transitivity.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This work was partially supported by Joint Funds of the National Natural Science Foundation of China (Grant No. U21B2020) and Ant Group Research Fund.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1, 2

[4] Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13877–13888, Singapore, 2023. Association for Computational Linguistics. 1, 2, 4

[5] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics, 2022. 2, 6, 7

[6] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32, 2019. 2

[7] Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Vlkeb: A large vision-language model knowledge editing benchmark, 2024. 1, 2

[8] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 4

[9] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. corr abs/2305.03726 (2023), 2023. 2

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2, 4

[11] Jiaqi Li, Miaozeng Du, Chuanyi Zhang, Yongrui Chen, Nan Hu, Guilin Qi, Haiyun Jiang, Siyuan Cheng, and Bozhong Tian. Mike: A new benchmark for fine-grained multimodal entity knowledge editing, 2024. 1, 2

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1, 2, 4

[13] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. 4

[14] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 1, 2, 4

[15] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1, 4

[16] Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022. 1, 4

[17] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 4

[18] Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. Updating language models with unstructured facts: Towards practical knowledge editing. *arXiv preprint arXiv:2402.18909*, 2024. 2

[19] Xunjian Yin, Jin Jiang, Liming Yang, and Xiaojun Wan. History matters: Temporal knowledge editing in large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19413–19421, 2024. 2

[20] Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, 2023. 1, 4

[21] Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore, 2023. Association for Computational Linguistics. 2, 4

[22] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 4