Supplementary Material for

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

1. Editing Methods and Multimodal Large Language Models

1.1. Knowledge Editing Methods

Fine-tune. Fine-tuning (FT) has become a widely adopted strategy for adapting pre-trained language models to specific tasks or domains. In previous benchmarks such as MMEdit [1], MIKE [7], and VLKEB [5], two primary fine-tuning methods were explored: one involves fine-tuning the final layer of multimodal language models, while the other focuses on fine-tuning the visual encoder. However, the latter approach has shown relatively poor performance. Given that our scenarios encompass both unimodal and multimodal editing, we opted to fine-tune only the final layer of the multimodal language model to effectively address all scenarios.

IKE. IKE [15] (In-Context Knowledge Editing) enables the modification and acquisition of new factual knowledge by embedding example demonstrations directly into the input data, eliminating the need for additional training phases. This approach allows for efficient updates to the system's knowledge base without altering model parameters.

MEND. MEND [10] employs lightweight model editor networks to modify the weights of a pre-trained model based on the fine-tuning gradient associated with a specific correction. By leveraging this gradient, MEND enhances the efficiency and precision of model edits.

SERAC. SERAC [11] introduces a memory-based model editing technique that utilizes an explicit memory system to store and retrieve edits. During inference, the memory is used to refine the base model's output. An auxiliary scope classifier determines whether the input falls within the relevant domain of the memory cache. If relevant, the input is combined with the appropriate memory item and processed by a counterfactual model to generate the final prediction.

ROME. ROME [9] identifies the specific layer within the Transformer architecture [13] where factual knowledge is stored, and modifies the feedforward network in that layer to incorporate the updated facts. This method ensures precise localization and efficient updating of the model's knowledge. For multimodal scenarios, where the subject s is represented as "the {entity} in the image," ROME has been tested only in unimodal knowledge editing scenarios due to its reliance on subject tokens being present in the input text.

1.2. Multimodal Large Language Models

BLIP2-OPT. BLIP2 [6] introduces an efficient and versatile pre-training framework that leverages frozen image encoders and large language models. It employs a lightweight Querying Transformer to effectively integrate vision and text modalities, achieving state-of-the-art performance across various vision-language tasks. For our experiments, we used BLIP2-OPT, which incorporates ViT-L [3] in the vision module and an unsupervised-trained OPT [14] model as the decoder-based language model.

MiniGPT-4. MiniGPT-4 is a robust vision-language model similar to BLIP2, utilizing a frozen visual encoder in conjunction with the frozen Vicuna [2] model. It introduces a projection layer to align visual features with the Vicuna language model. Like BLIP2, MiniGPT-4 uses a ViT-G/14 from EVA-CLIP [12] and a Q-Former as part of its visual processing component.

*Equal contribution. † Corresponding author.

LLaVA-1.5. LLaVA [8] enhances its ability to handle complex multimodal tasks by aligning different modalities during pre-training and refining response generation through instruction-based fine-tuning. By pre-training and fine-tuning an alignment network with Vicuna, LLaVA significantly improves its performance in managing intricate multimodal interactions.

1.3. Metrics

The effectiveness of knowledge editing was evaluated using the metrics defined in Section 3. For unimodal knowledge editing, we applied the five core metrics: **Reliability**, **Generality**, **Portability**, **Locality**, and **Stability**. In transitive and multimodal scenarios, where test data included multimodal elements, we assessed the model's performance across different images. In these cases, Reliability was subdivided into $\textbf{Reliability}_{\textbf{S}}$ and $\textbf{Reliability}_{\textbf{U}}$, where $\textbf{Reliability}_{\textbf{S}}$ evaluates images seen during the editing process, and $\textbf{Reliability}_{\textbf{U}}$ assesses unseen images. Generality was similarly divided into $\textbf{Generality}_{\textbf{S}}$ and $\textbf{Generality}_{\textbf{U}}$. Locality was split into $\textbf{Locality}_{\textbf{T}}$ for textual data and $\textbf{Locality}_{\textbf{I}}$ for multimodal data. Portability and Stability were evaluated as in unimodal editing.

2. Complete experimental results

We present the results of LLava-1.4 experiments in Section **??**, and the complete experimental results are in Table 1, 2 and 3.

3. Prompt for Dataset Construction

During the dataset construction process, we employed GPT-4o to extract entity types from the questions. GPT-4o was further used to paraphrase questions in order to build the Generality dataset. Additionally, based on the questions and extracted entity types, GPT-4o provided examples of the same entity types to construct the Stability dataset. The specific prompts provided to GPT-4o are illustrated in Table 7, 8 and 9.

4. Experiment Details

In our experiments, we employed the single editing approach [4], which updates one piece of knowledge at a time and then evaluates the results. In single editing, memory-based methods benefit from storing just one new piece of knowledge, while parameter-update methods can effectively integrate this single update.

For experiments mentioned in this paper, we use eight Nvidia A100 GPUs with 80GB memory. And detailed parameters are listed in Table 4, 5 and 6.

References

[1] 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] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, pages 47934–47959. Curran Associates, Inc., 2023. 2

[5] 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

[6] 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

[7] 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

[8] 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. 2

[9] 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

[10] 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

[11] 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

[12] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[13] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

Model	Method	Reliability	Generality	Portability	Locality	Stability
BLIP2-OPT (3.8B)	FT	100.0	100.0	15.7	82.6	28.5
	IKE	100.0	98.5	36.3	54.3	46.9
	MEND	97.4	96.8	22.6	99.3	37.8
	SERAC	100.0	100.0	38.1	100.0	14.7
	ROME	90.8	90.1	23.6	95.9	59.9
MiniGPT-4 (7.8B)	FT	100.0	100.0	24.2	90.1	38.4
	IKE	99.7	98.2	41.7	53.5	50.2
	MEND	99.8	99.7	28.5	99.4	51.7
	SERAC	100.0	100.0	38.0	100	14.9
	ROME	98.4	97.5	30.9	97.2	53.2
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 different MLLMs. For each model, 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
BLIP2-OPT (3.8B)	FT	99.3	99.3	99.2	99.3	14.7	82.6	48.3	28.5
	IKE	95.3	94.9	96.6	96.2	28.9	54.3	2.4	31.2
	MEND	60.3	59.6	59.7	59.6	13.6	99.1	73.0	33.2
	SERAC	78.0	81.9	80.9	81.9	11.7	100.0	2.9	11.0
	ROME	27.0	27.0	24.5	24.6	11.2	97.3	66.6	39.7
MiniGPT-4 (7.8B)	FT	99.7	99.8	99.6	99.7	24.5	90.1	48.1	45.0
	IKE	99.1	99.3	95.1	95.1	38.3	53.5	3.2	42.3
	MEND	65.5	66.4	64.7	65.1	23.6	99.3	89.8	48.0
	SERAC	52.1	52.8	46.2	48.0	0.8	99.8	3.5	0.9
	ROME	29.8	29.7	28.9	29.1	20.6	96.7	87.2	49.5
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 different MLLMs. For each model, the best results are indicated in a dark color, and the second-best results are indicated in a light color.

[14] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1

[15] 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

Model	Method	$\text{Reliability}_\text{S}$	$\text{Reliability}_\text{U}$	$\text{Generality}_\text{S}$	$\text{Generality}_\text{U}$	Portability	Locality_T	Locality_I	Stability
BLIP2-OPT (3.8B)	FT	100.0	100.0	100.0	98.1	22.2	91.6	45.8	7.9
	IKE	99.3	98.2	98.9	96.2	29.8	57.5	2.5	29.6
	MEND	94.3	92.2	93.7	85.5	16.3	98.8	69.7	10.6
	SERAC	79.8	80.0	79.7	70.1	30.2	100.0	2.8	2.3
MiniGPT-4 (7.8B)	FT	100.0	100.0	100.0	98.9	22.2	91.6	45.8	7.9
	IKE	100.0	100.0	99.4	99.8	44.9	55.0	3.5	40.1
	MEND	95.5	94.7	96.8	84.7	25.8	98.3	52.8	16.7
	SERAC	59.5	60.2	55.1	45.9	1.1	97.9	3.5	0.0
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 different MLLMs. For each model, the best results are indicated in a dark color, and the second-best results are indicated in a light color.

Models	MaxIter	Edit Num	Optimizer	LR
BLIP2-OPT	20000	1	Adam	1e-6
MiniGPT-4	20000	1	Adam	1e-6
LLaVA-1.5	20000	1	Adam	1e-6

Table 4. FT hyper-parameters

Models	MaxIter	Edit Num	Optimizer	LR
BLIP2-OPT	30000	1	Adam	1e-6
MiniGPT-4	30000	1	Adam	1e-6
LLaVA-1.5	30000	1	Adam	1e-6

Table 5. MEND hyper-parameters

Models	MaxIter	Edit Num	Optimizer	LR
BLIP2-OPT	30000	1	Adam	1e-5
MiniGPT-4	30000	1	Adam	1e-5
LLaVA-1.5	30000	1	Adam	1e-5

Table 6. SERAC hyper-parameters

System:
You will be given a question and the subject of the question, you need to give the entity type of the subject. Here is an example:
Example:
Subject: George Rankin
Question: What is George Rankin's occupation?
Entity Type: person

User:
Subject: The French Lieutenant's Woman
Question: Who is the author of The French Lieutenant's Woman?

Table 7. Prompt for entity type extraction.

System:
You will be given a question and the subject of the question, you need to come up with a semantically similar paraphrase question. Here is an example:
Example:
Subject: George Rankin
Question: What is George Rankin's occupation?
Paraphrase Question: What does George Rankin do for a living?

User:
Subject: The French Lieutenant's Woman
Question: Who is the author of The French Lieutenant's Woman?

Table 8. Prompt for Generality dataset construction.

System:
You will be given a question and the subject of the question, you need to give a subject of the same entity type but different and give the answer that this subject should have answered under the original question. Here is an example:
Example:
Subject: George Rankin
Question: What is George Rankin's occupation?
Subject of same entity type: Leo Messi
Answer: Professional footballer

User:
Subject: The French Lieutenant's Woman
Question: Who is the author of The French Lieutenant's Woman?

Table 9. Prompt for Stability dataset construction.