Visual-Oriented Fine-Grained Knowledge Editing for MultiModal Large Language Models

Supplementary Material

A. Experiments on more base models

We conducted experiments based on LLaVA-NeXT, the results are shown in Table 5.

	Specificity	Reliability	Locality	Generality
FT-LLM	21.82	84.0	82.91	84.0
FT-Visual	19.32	85.71	100.0	84.63
IKE	24.56	99.93	56.80	85.35
SERAC	31.42	98.60	100.0	95.96
MSCKE	52.23(↑20.81)	99.92(↑1.32)	100.0(↑0.00)	97.16(↑1.20)
MEND	65.43	96.90	96.87	96.54
MSCKE-MEND	67.86(↑2.43)	97.60(↑0.70)	100.0(↑3.13)	96.58(↑0.04)

Table 5. Experiments on LLaVA-NeXT-7b.

B. Analysis on Similarity Threshold

In Eq. 3, we set the similarity threshold between the base model and the counterfactual model to 0.5. To evaluate the effect of this threshold on classifier accuracy, we conducted an experiment, with the results presented in Fig. 3. The experimental results show that our classifier is insensitive to the threshold, and only when the threshold is set too high or too low will the classifier performance significantly decrease.

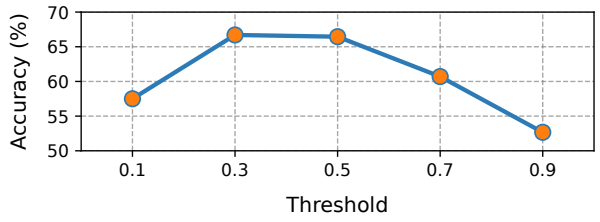


Figure 6. Classifier accuracy under different thresholds.

C. Computational Cost Analysis

The MSCKE framework comprises a multimodal scope classifier, a base model, and a counterfactual model. We evaluate the computational cost of these components in terms of inference time and memory usage. The results on BLIP-2 OPT are shown in Table 6.

Our analysis shows that the multimodal scope classifier requires significantly less inference time and memory compared to the base model. These findings highlight that the classifier introduces minimal computational overhead while playing a crucial role in improving editing performance. Its lightweight design ensures that it does not become a bottleneck, making the MSCKE framework well-suited for real-world applications where computational resources may be limited.

	inference time	parameter size
classifier	36ms	427.62M
base	121ms	3.8B
counterfactual	85ms	1.2B

Table 6. Comparison of inference time and parameter size among the multimodal scope classifier, base model, and counterfactual model on BLIP-2 OPT.

Components	CLIP-ViT-B/32	CLIP-ViT-L/14
concatenation	63.70	63.80
cross-attention	64.45	64.35
dot-product attention	**64.73**	**64.85**

Table 7. Implementation analysis of multimodal scope classifier. The accuracy of the classifier is displayed. Using CLIP-ViT-B/32 for feature extraction, coupled with dot-product attention for feature fusion, yields strong performance.

D. Implementation Analysis of Multimodal Scope Classifier

The multimodal scope classifier consists of two key components: a feature extraction module and a feature fusion module. We investigate how different implementations of these modules impact the classifier's performance. The feature extraction module can be implemented using either CLIP-ViT-B/32 or CLIP-ViT-L/14, while the feature fusion module can be realized through feature concatenation, cross-attention, or dot-product attention. The results of scope classifier on FGVEdit dataset are shown in Table 7. Regarding feature fusion, dot-product attention delivers the best performance with minimal computational overhead. Its superior efficiency in capturing essential interactions between text and image features, without the complexity of methods like cross-attention, makes it an ideal choice for feature fusion in our multimodal classifier.

For feature extraction, CLIP-ViT-B/32 performs similarly to CLIP-ViT-L/14, suggesting that CLIP-ViT-B/32 is sufficient for extracting comprehensive features. Additionally, we selected the more powerful SigLIP as the feature extractor for experimentation. The results are presented in

	BLIP-2 OPT				MiniGPT-4			
	Specificity	Reliability	Locality	Generality	Specificity	Reliability	Locality	Generality
MSCKE(clip)	61.60	99.13	100.0	98.56	57.20	99.50	100.0	93.00
MSCKE(siglip)	61.71	99.71	100.0	99.11	57.38	99.68	100.0	93.56
MSCKE-MEND(clip)	68.38	97.40	100.0	96.50	71.98	97.05	100.0	96.70
MSCKE-MEND(siglip)	68.42	97.95	100.0	96.72	72.21	97.65	100.0	97.21

Table 8. Experiments on siglip.

	Method	EDITING VQA				EDITING IMAGE CAPTION			
		Reliability ↑	T-Generality ↑	T-Locality ↑	M-Locality ↑	Reliability ↑	T-Generality ↑	T-Locality ↑	M-Locality ↑
	BLIP-2 OPT								Size:3.8B
Base Methods	Base Model	0.00	0.00	100.0	100.0	0.00	0.00	100.0	100.0
	FT(vision block)	60.56	49.79	100.0	8.47	18.94	5.86	100.0	8.40
	FT(last layer)	57.66	46.70	21.67	3.06	16.60	3.50	24.96	7.12
Model Editing	Knowledge Editor	85.28	84.23	90.31	52.48	0.30	0.10	88.31	49.52
	In-Context Editing	99.71	91.59	48.79	2.53	83.80	69.40	48.95	2.95
	SERAC	99.90	99.90	100.0	2.91	98.90	98.90	99.98	7.52
	MSCKE	99.80	95.60	100.0	98.08	97.20	97.40	100.0	99.34
	MEND	98.51	97.51	99.94	96.65	80.00	78.10	94.54	70.84
	MSCKE-MEND	98.20	92.10	100.0	99.23	80.10	78.70	100.0	99.33
	MiniGPT-4								Size:7.3B
Base Methods	Base Model	0.00	0.00	100.0	100.0	0.00	0.00	100.0	100.0
	FT(vision block)	36.3	0.3	100.0	9.29	3.10	0.00	100.0	8.56
	FT(last layer)	0.10	0.00	72.60	15.75	0.00	0.00	53.50	12.68
Model Editing	Knowledge Editor	95.37	92.64	97.31	73.76	75.50	67.80	97.15	69.92
	In-Context Editing	100.0	94.40	50.30	3.67	77.00	51.80	52.18	4.68
	SERAC	99.90	92.60	99.90	5.52	97.30	74.60	99.89	7.20
	MSCKE	99.90	92.60	100.0	84.12	98.60	88.90	100.0	99.80
	MEND	96.20	95.40	98.23	81.08	80.80	78.60	98.41	75.25
	MSCKE-MEND	97.40	92.50	100,0	99.74	82.0	80.70	100.0	99.83

Table 9. Main results on the MMEdit. T-Locality, M-Locality refer to the textual and multimodal stability. T-Generality represents textual generality. Reliability denotes the Exact Match of successful editing.

Table 8. Experiments demonstrate that the stronger feature extractor did not lead to significant performance improvements. This finding highlights that our classifier does not require the more computationally expensive CLIP-ViT-L/14, making the approach more efficient and suitable for practical applications with limited computational resources.

E. Experimental Results on Other Benchmark

We also conducted experiments on MMEdit, and the results are shown in Table 9. All baseline results come from MMEdit. The findings demonstrate that our method achieves strong performance on MMEdit as well.

F. Dataset Examples

More examples of our dataset are shown in Fig. 7 and Fig. 8.

Editing Sample
Q: What is the name of the street?
A: 34th street
Out-of-Visual-Scope
Q: Which side of the sign is the road on?
A: left
In-Visual-Scope
Q: What do the two white signs say?
A: 34th street
Generality
Q: Can you tell me the name of the street?
A: 34th street
Locality
Q: who sings the song middle finger in the air?
A: Cobra Starship

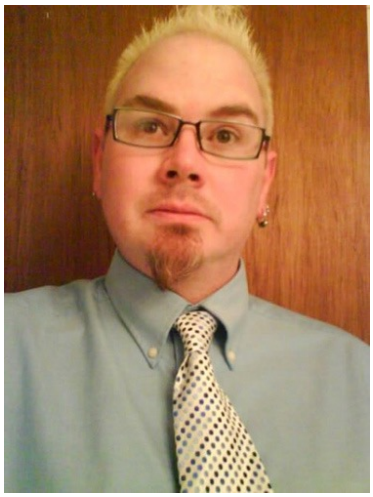
Editing Sample
Q: What is the color of the hydrant?
A: red
Out-of-Visual-Scope
Q: How many different colors does the cone have?
A: 2
In-Visual-Scope
Q: Is the fire hydrant the same color as Christmas?
A: yes
Generality
Q: What hue is the fire hydrant?
A: red
Locality
Q: who wrote and sang ground control to major tom?
A: David Bowie

Editing Sample
Q: Is there water in the picture?
A: yes
Out-of-Visual-Scope
Q: Is the sky clear?
A: yes
In-Visual-Scope
Q: What is in front of the buildings?
A: water
Generality
Q: Does the image contain any water?
A: yes
Locality
Q: when did the development of nuclear science begin
A: 1898

Figure 7. Some examples of our dataset.

Editing Sample
Q: What is the sign seen?
A: stop
Out-of-Visual-Scope
Q: What shape is the majority?
A: square
In-Visual-Scope
Q: What does it say above stop?
A: notice
Generality
Q: What is the displayed sign?
A: stop
Locality
Q: where are kia's made in the us?
A: West Point, Georgia

Editing Sample
Q: How many urinals can you see?
A: 1
Out-of-Visual-Scope
Q: Is this room well lit?
A: no
In-Visual-Scope
Q: Which of these objects is a urinal?
A: right
Generality
Q: What is the total number of urinals visible?
A: 1
Locality
Q: where did clay matthews jr play college football?
A: University of Southern California



Editing Sample
Q: Is this man wearing a necktie?
A: yes
Out-of-Visual-Scope
Q: How many earrings does the guy have in his ear?
A: 1
In-Visual-Scope
Q: How many dots on the necktie?
A: 50
Generality
Q: Does this man have a necktie on?
A: yes
Locality
Q: when do casey and cappie get back together?
A: Legacy

Figure 8. Some examples of our dataset.