VLKEB: A Large Vision-Language Model Knowledge Editing Benchmark

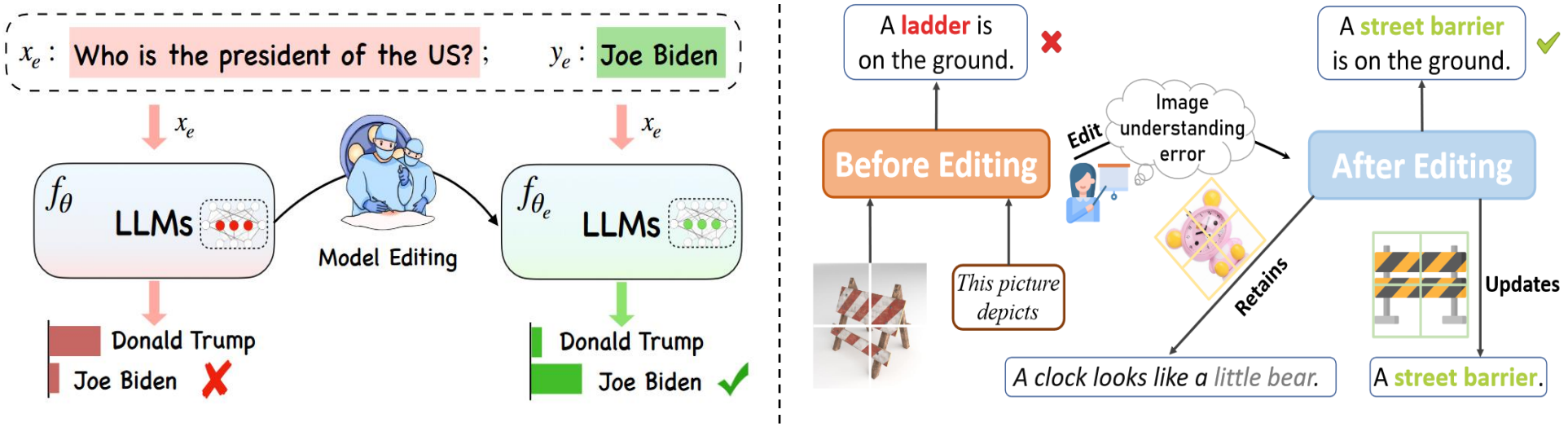
Han Huang*, Haitian Zhong*, Tao Yu,

Qiang Liu†, Shu Wu, Liang Wang, Tieniu Tan

New Laboratory of Pattern Recognition (NLPR), CASIA

Knowledge editing in LLM and MLLM

- Why do we need knowledge editing?
  - The model has learned knowledge that is undesirable, such as outdated knowledge, biases, privacy-sensitive data, etc.
- Problem definition:
  - Efficiently change the model's output for specific knowledge to the desired response without retraining the entire model and affecting results for other inputs.



Left Image: *Editing Large Language Models: Problems, Methods, and Opportunities*. Yunzhi Yao et al.
Right Image: *Can We Edit Multimodal Large Language Models?* Siyuan Cheng et al.

Four Metrics of Knowledge Editing

Reliability:

- Success rate in producing the correct output for the edited content.

Generality:

- Generalize within the scope of the edit, for instance, in handling rephrased questions or similar images under the same entity.

Locality:

- Ensuring the editing succeeds without impacting unrelated knowledge.

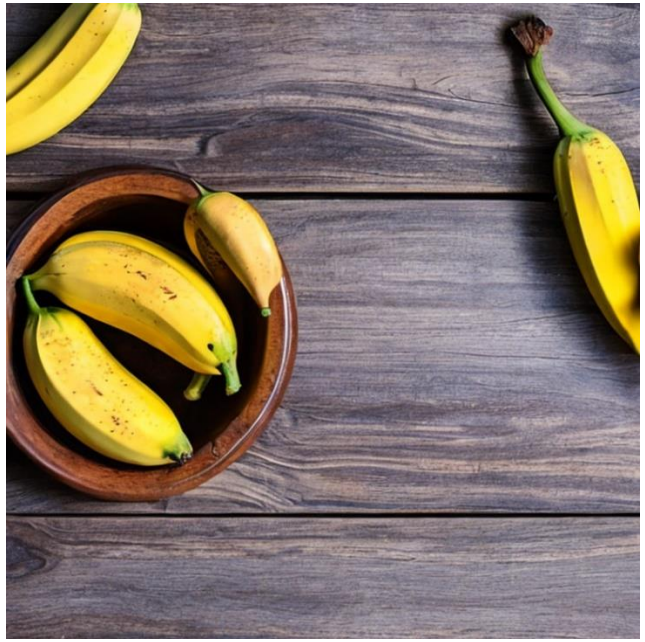
Portability:

- Ability to apply edited knowledge to questions associated with that knowledge.

Motivation

Limitations of existing benchmark:

- Synthetic images in testing do not match the questions and answers
- Lack of portability evaluation

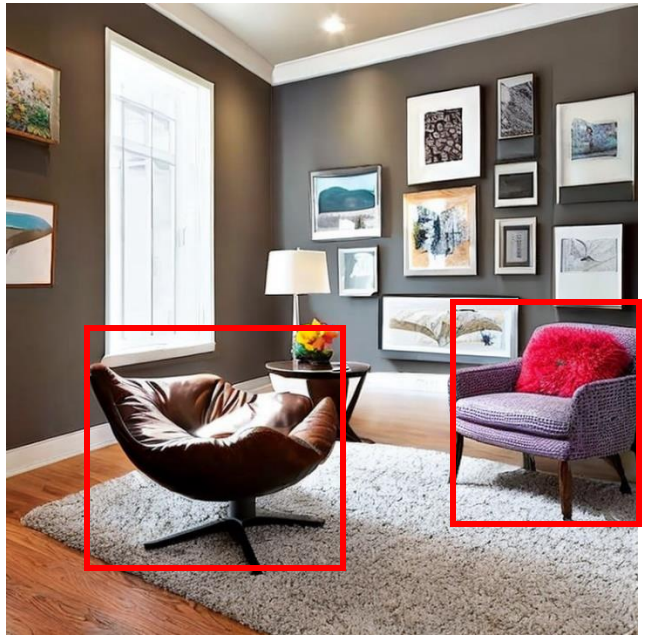


Synthetic image

Q: How many fruits are on the plate?

A: 7

Original image

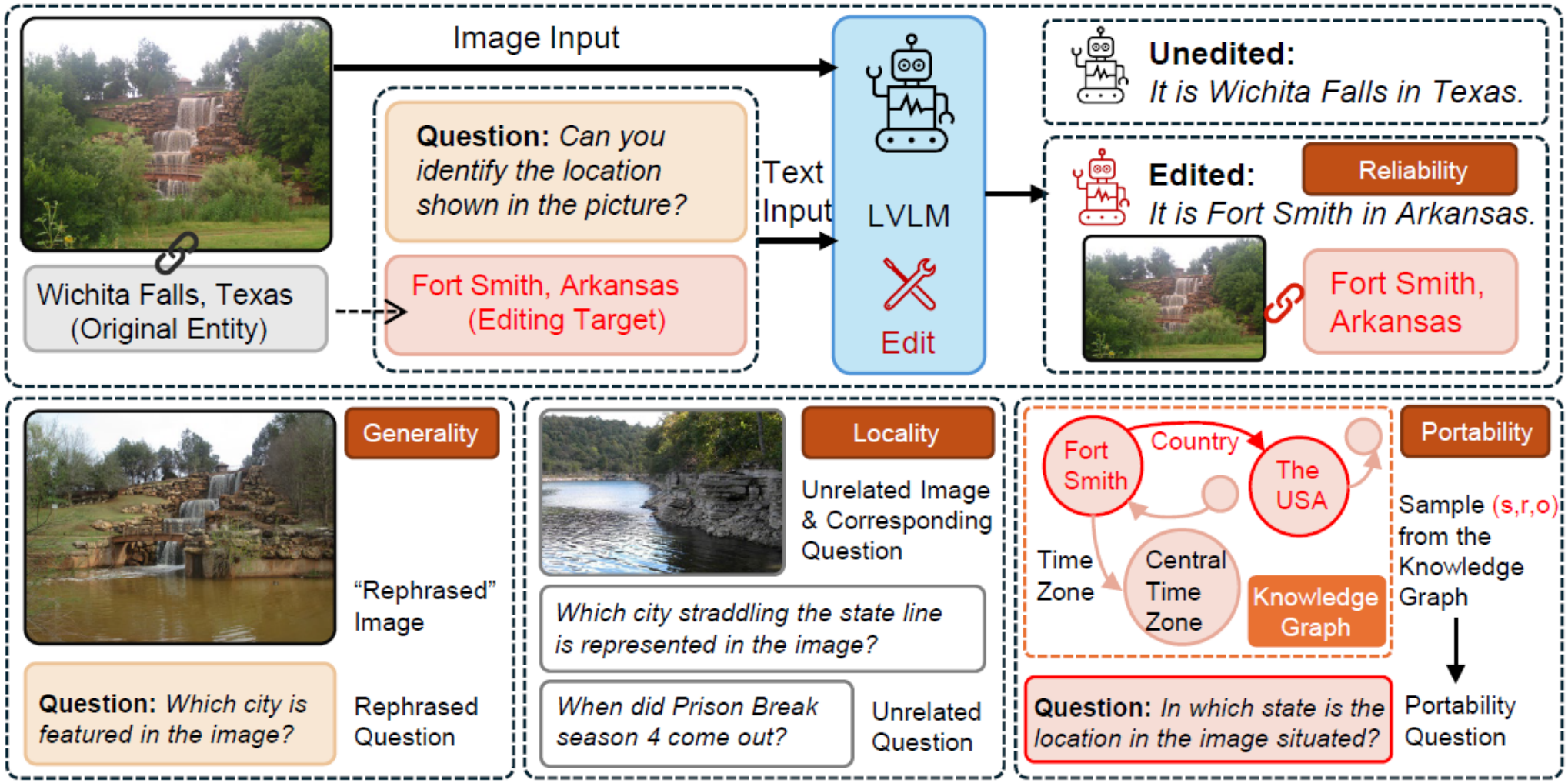


Synthetic image

a photo of ___

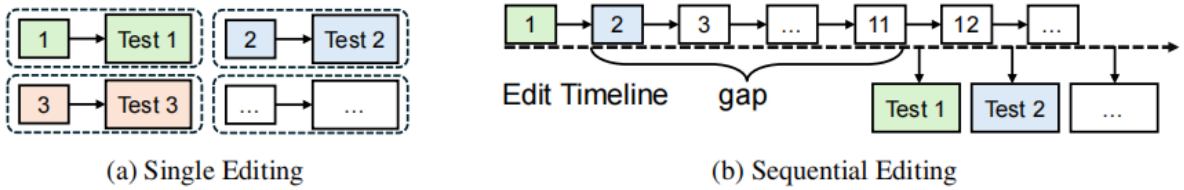
white ornate seat in nicely decorated room with television.

Construction of VLKEB



Experiment Setup

- Single editing and Sequential editing



(a) Single Editing (b) Sequential Editing

- Knowledge Editing Methods
  - Keep original parameter: IKE, SERAC
  - Update model parameter: FT, MEND, KE

- Tested LVLMs

Table 6: LVLM versions in experiments. (Vis.: Vision Encoder)

LVLM	**BLIP2-OPT**	**MiniGPT-4**	**LLaVA-1.5**	**mPLUG-Owl2**	**Qwen-VL**
LLM	OPT-2.7B	Vicuna-7B	Vicuna-7B-v1.5	LLaMA-2-7B	Qwen-7B
Vis.	ViT-g (1B)	ViT-g (1B)	ViT-L (0.3B)	ViT-L (0.3B)	ViT-G (1.9B)

Experiment Results - Single Editing

Reliability and generality are both high

- Memory-based method stores only one piece of new knowledge
- The parameter update methods adapt well to single pieces of new knowledge

Differences in textual and visual locality tests

- Memory-based method has stronger impact on visual locality
- The parameter update methods also influence the expression of unrelated knowledge

Portability is generally poor

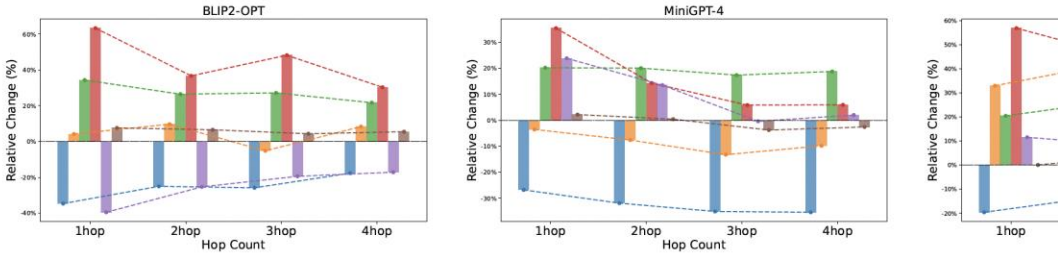
- The effect of current knowledge editing is generally unsatisfactory
- The model struggles to effectively utilize edited knowledge

Table 2: The single editing results of various editing methods applied to different LVLMs. (Rel.: Reliability; T/I-Gen.: Text/Image Generality; T/I-Loc.: Text/Image Locality; Port.: Portability)

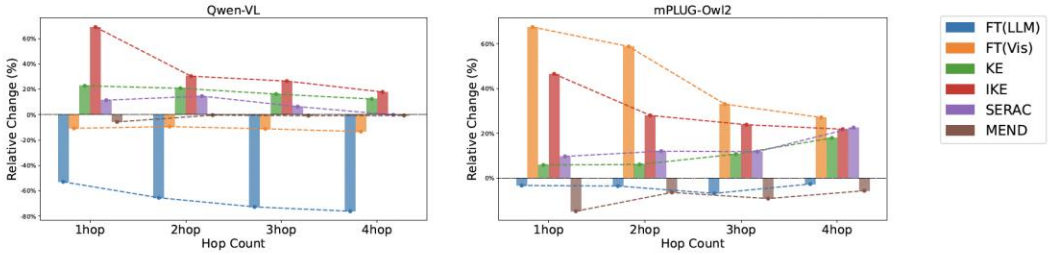
Model	Method	Rel.↑	T-Gen.↑	I-Gen.↑	T-Loc.↑	I-Loc.↑	Port.↑
BLIP2-OPT (∼ 3.8 B)	FT (LLM)	**99.75**	99.08	98.95	71.10	19.90	17.13
	FT (Vis)	99.33	96.68	99.13	99.99	5.30	27.22
	KE	94.45	92.40	93.34	64.13	12.22	34.73
	IKE	99.47	**99.40**	**99.56**	70.11	10.26	**44.22**
	SERAC	96.02	95.99	96.01	**100.0**	2.40	15.25
	MEND	98.52	98.42	98.47	99.34	**89.05**	28.80
MiniGPT-4 (∼ 7.8 B)	FT (LLM)	99.60	98.72	98.10	90.17	35.39	27.13
	FT (Vis)	**100.0**	84.89	99.19	**99.99**	20.26	37.06
	KE	98.47	97.89	98.11	75.47	16.14	48.06
	IKE	99.98	**99.68**	**99.98**	59.25	9.73	**54.30**
	SERAC	99.37	97.30	99.29	99.93	4.54	49.22
	MEND	99.20	98.98	99.15	99.46	**92.67**	40.09
LLaVA-1.5 (∼ 7 B)	FT (LLM)	99.59	99.43	99.31	86.34	29.24	30.23
	FT (Vis)	99.80	99.12	97.55	**99.99**	18.79	54.43
	KE	99.07	97.59	98.65	77.36	15.25	48.62
	IKE	**99.99**	99.66	**100.0**	68.65	14.25	**63.33**
	SERAC	99.93	**99.78**	99.93	99.98	1.91	45.03
	MEND	99.54	99.21	99.52	99.36	**90.14**	40.39
Qwen-VL (∼ 9.7 B)	FT (LLM)	97.92	96.30	95.48	72.80	37.23	16.15
	FT (Vis)	**100.0**	95.27	62.28	**100.0**	14.14	30.61
	KE	98.71	98.70	98.26	72.09	52.63	42.10
	IKE	99.01	98.85	**99.01**	57.97	10.48	**57.99**
	SERAC	97.62	95.68	97.84	99.85	0.81	38.22
	MEND	99.54	**99.36**	97.76	97.75	**78.65**	32.35
mPLUG-Owl2 (∼ 8.2 B)	FT (LLM)	99.21	95.72	99.39	71.42	34.31	42.77
	FT (Vis)	97.24	96.36	82.39	**99.99**	50.14	**74.09**
	KE	89.10	88.37	88.62	55.80	21.07	46.82
	IKE	**99.98**	**99.93**	**100.0**	64.88	16.59	64.83
	SERAC	99.03	97.73	98.99	99.97	1.32	48.52
	MEND	98.65	98.15	94.26	99.56	**90.47**	37.68

Experiment Results – Portability

- Multi-hop Portability: Performance decreases as the number of hops increases



(a) BLIP2-OPT (b) MiniGPT-4 (c) LLaVA-1.5



(d) Qwen-VL (e) mPLUG-Owl2

Table 3: Portability increases after additionally edit corresponding one-hop knowledge.

Portability	FT (LLM)	FT (Vis)	SERAC	MEND
Blip2-OPT	17.13→46.71 (↑ 29.58)	27.22→39.00 (↑ 11.78)	16.16→32.12 (↑ 15.96)	28.75→68.18 (↑ 39.43)
MiniGPT-4	27.13→44.65 (↑ 17.52)	37.06→56.91 (↑ 19.85)	47.49→51.10 (↑ 3.61)	39.19→53.83 (↑ 14.64)
LLaVA-1.5	30.23→74.79 (↑ 44.56)	54.43→84.81 (↑ 30.38)	45.03→72.75 (↑ 27.72)	40.39→70.73 (↑ 30.34)
Qwen-VL	16.15→88.88 (↑ 72.73)	30.61→55.15 (↑ 24.54)	38.22→54.44 (↑ 16.22)	32.35→69.41 (↑ 37.06)
mPLUG-Owl2	42.77→59.37 (↑ 16.60)	74.09→93.80 (↑ 19.71)	48.52→67.91 (↑ 19.39)	37.68→70.71 (↑ 33.03)

Figure 3: Relative change (compared with unedited base model) of Multi-hop Portability results.

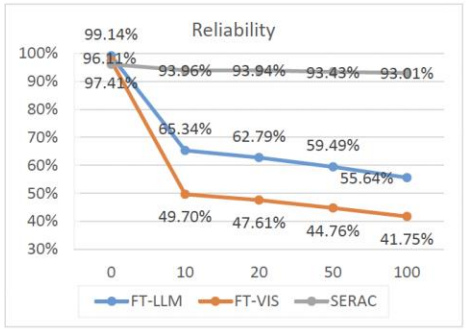
- Adding one-hop knowledge edits leads to improvements across the board.

  New effective methods are still needed.

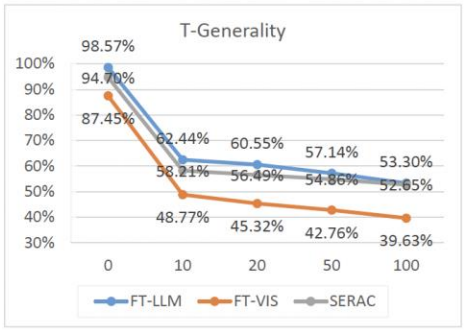
Experiment Results – Sequential Editing

- Some editing methods are inapplicable (IKE, KE)
- Certain parameter update methods cause model collapse (MEND)
- In memory-based methods, as the amount of stored memory grows, performance declines due to limitations in retrieval methods.
- The parameter update methods result in forgetting and degrade model performance.

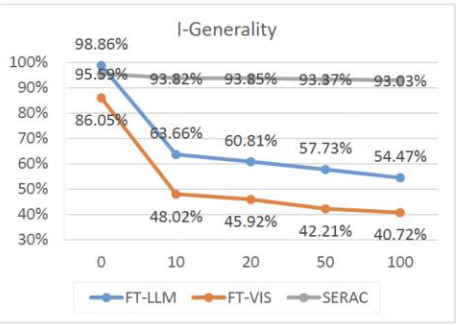
Current editing methods have limitations



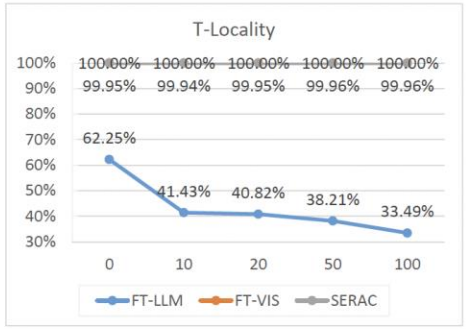
(a) Average Reliability



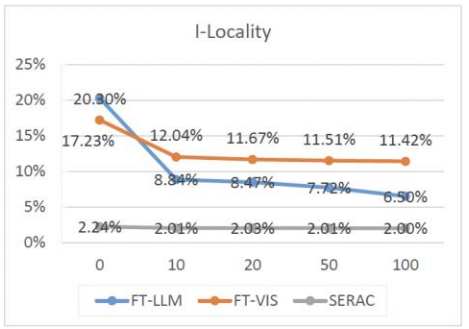
(b) Average Text Generality



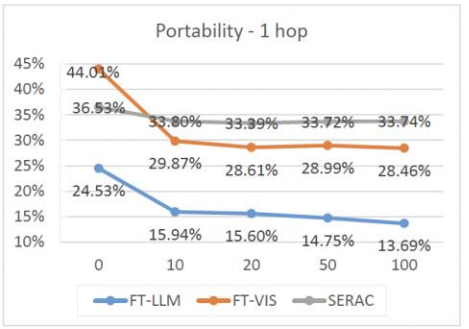
(c) Average Image Generality



(d) Average Text Locality



(e) Average Image Locality



(f) Average Portability 1-hop

Future Directions

Direct LVLM Editing

- These methods are adapted from LLM and are not specifically designed for LVLMs. Research could explore direct LVLM editing methods to address this gap.

Sequential Editing in LVLM

- We have observed performance degradation across methods in sequential editing, future work on LVLM editing should focus on mitigating these issues.

Portability Evaluation

- Future research should further emphasize Portability as an important aspect.