ROSE: Risk-Aware Orthogonal Subspace Navigation for Lifelong Knowledge Editing in Multimodal Large Language Models

Lingyun Song[1,2*] Ziyao Chen[1] Kang Pan[3] Xiaolin Han[1*] Yifei Zhang[1] Xiaoqi Wang[1] Yudai Pan[1*] Xuequn Shang[1,4*]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an [2] Shenzhen Research Institute of Northwestern Polytechnical University, Shenzhen

[3] Independent Researcher [4] Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an

Code — https://github.com/sunlitsong/ROSE

Introduction & Motivation

- **Outdated Knowledge:** MLLMs possess knowledge that can become outdated over time and require **continuous factual updates** to adapt to an evolving world.
- **Catastrophic Forgetting:** Sequential edits **interfere with previously edited knowledge** due to overlapping parameter subspaces and cross-modal entanglement.
- **Locality Erosion:** Existing editors struggle to precisely activate relevant edited knowledge, causing **unintended interference** with unrelated queries and pre-trained capabilities.

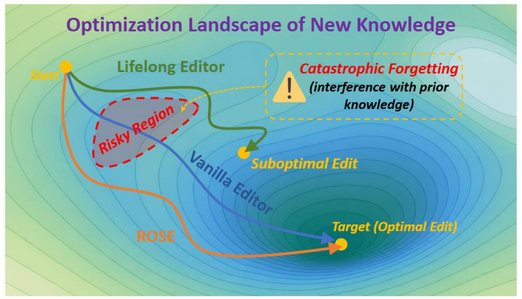


Figure 1. ROSE navigates editing trajectories away from high-risk subspaces.

Overview & Contributions

- **Risk-Guided Orthogonal Editing:** Navigates parameter updates into **orthogonal subspaces** and dynamically **protects high-risk parameters** to mitigate catastrophic forgetting.
- **Multi-Granularity Knowledge Integration:** Combines **prototype-anchored semantic gating** with **perturbation-driven subspace fusion** to precisely activate edited knowledge and mitigate locality erosion.
- **Superior Lifelong Editing Performance:** Achieves competitive performance across **two multimodal benchmarks** while maintaining remarkable stability over 1,000 sequential edits.

Experimental Results

Table 1. Performance comparison across multimodal benchmarks.

Edits	Methods	VLKEB					E-IC					E-VQA				
		Reliability	Generality		Locality		Reliability	Generality		Locality		Reliability	Generality		Locality	
			T-Gen.	M-Gen.	T-Loc.	M-Loc.		T-Gen.	M-Gen.	T-Loc.	M-Loc.		T-Gen.	M-Gen.	T-Loc.	M-Loc.
T=1	FT-L	94.29	87.00	92.22	91.16	91.37	73.48	72.98	65.79	99.28	99.06	93.88	87.98	80.25	99.61	94.78
	FT-M	76.31	65.57	59.43	100.00	92.35	56.19	56.55	49.94	100.00	100.00	87.29	76.11	53.23	100.00	96.95
	MEND	92.13	91.28	90.22	89.19	90.13	92.82	91.81	90.59	96.38	93.69	91.23	90.05	91.29	91.02	90.22
	SERAC	89.77	89.11	87.92	66.68	14.20	88.18	81.03	85.61	84.01	28.58	89.33	83.72	84.97	82.05	23.78
	LEMoE	94.85	93.09	91.67	87.03	87.88	93.80	91.42	90.61	95.14	93.00	93.60	92.77	89.99	99.28	96.98
	RECIPE	92.67	92.35	91.01	89.67	82.85	84.45	76.97	81.57	96.53	96.37	91.37	86.51	87.73	94.27	88.88
	LiveEdit	96.43	95.22	**93.72**	100.00	**100.00**	82.16	81.01	78.27	100.00	**100.00**	94.28	**94.51**	88.01	100.00	100.00
	ROSE	**100.00**	**98.85**	88.87	**100.00**	98.72	**100.00**	**94.23**	**94.90**	**100.00**	99.55	**100.00**	93.92	**92.92**	**100.00**	**100.00**
T=100	FT-L	75.41	73.67	74.16	70.01	82.05	65.08	60.90	58.46	86.82	89.19	79.67	70.05	64.07	83.47	54.44
	FT-M	60.60	59.79	56.29	100.00	68.07	59.41	55.44	52.16	100.00	72.93	82.90	72.83	47.26	100.00	43.39
	MEND	0.56	0.58	0.66	0.18	0.07	56.83	56.89	53.07	87.63	84.64	2.22	2.20	2.21	0.21	0.62
	SERAC	72.25	62.43	70.68	53.73	13.69	53.35	53.70	49.41	48.04	17.25	88.08	81.53	82.48	62.13	12.90
	LEMoE	83.07	75.36	71.72	54.09	49.68	55.36	53.14	51.12	87.77	81.35	42.41	36.60	34.33	78.57	53.28
	RECIPE	73.97	63.73	72.69	86.2	82.59	53.23	53.75	49.36	87.64	95.52	89.86	83.32	84.82	87.37	85.08
	LiveEdit	94.56	90.65	89.56	100.00	**100.00**	80.81	78.77	63.52	100.00	100.00	93.54	**92.34**	**85.89**	100.00	99.31
	ROSE	**98.22**	**95.00**	**93.80**	**100.00**	95.31	**93.28**	**87.55**	**82.30**	**100.00**	**100.00**	**97.00**	85.22	83.97	**100.00**	**99.38**
T=1000	FT-L	68.14	66.38	66.98	65.61	75.35	59.78	54.99	54.17	65.37	78.96	71.39	59.83	57.41	55.55	48.99
	FT-M	53.41	48.80	43.16	100.00	57.03	49.21	47.75	43.81	100.00	35.14	69.57	56.34	44.07	100.00	41.47
	MEND	0.03	0.05	0.07	0.06	0.08	54.39	54.14	50.99	83.87	80.60	0.04	0.05	0.05	0.08	0.09
	SERAC	60.93	56.49	60.06	52.94	15.04	52.93	53.44	49.01	49.91	16.65	85.57	75.58	82.01	62.46	15.69
	LEMoE	67.97	61.07	58.16	48.48	44.06	34.50	31.38	28.14	82.09	75.88	30.80	25.75	24.32	71.45	46.23
	RECIPE	62.00	56.84	61.50	85.37	82.07	53.11	53.48	48.99	87.93	94.84	87.00	76.81	83.09	86.95	87.03
	LiveEdit	92.22	83.97	82.75	100.00	**100.00**	72.80	69.95	57.05	100.00	**99.79**	92.93	**90.16**	**84.30**	100.00	**96.43**
	ROSE	**95.24**	**92.74**	**84.53**	**100.00**	88.39	**90.65**	**82.52**	**72.22**	**100.00**	97.42	**97.59**	80.85	77.81	**100.00**	95.52

Table 2. Ablation study

Settings	Reliability	Generality		Locality		Average
		T-Gen	M-Gen.	T-Loc.	M-Loc.	
Fine Tuning	79.16	76.75	75.94	14.54	7.44	50.77
w/o Multimodal Decoupling	91.05	88.14	85.63	100.00	40.54	81.07
w/o Knowledge Integration	93.15	90.81	90.05	57.38	15.83	69.44
w/o Semantic Gating	91.98	89.71	89.02	76.87	25.97	74.71
w/o Knowledge Fusion	96.28	93.90	83.30	100.00	84.37	91.57
w/o Orthogonal Editing	78.71	76.00	71.05	100.00	85.50	82.25
w/o Risk Awareness	90.07	87.31	78.09	100.00	88.14	88.72
ROSE	95.24	92.74	84.53	100.00	88.39	92.18

Table 3. Computational Efficiency

Edits	Editing Time (s/edit)	Inference Time (s/query)	Peak VRAM (MB)
1	1.6748	0.9973	16781.52
10	1.7113	1.1849	16826.89
100	1.3975	1.1248	16857.73
1000	1.3140	1.1256	17005.69

Figure 3. Semantic Gating Analysis

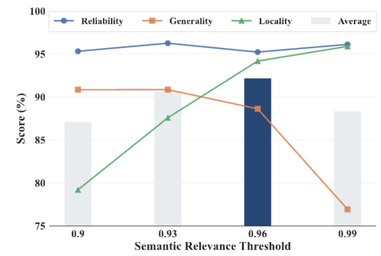
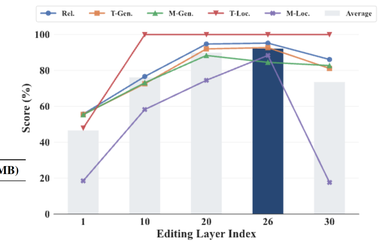


Figure 4. Editing Depth Analysis



Methodology

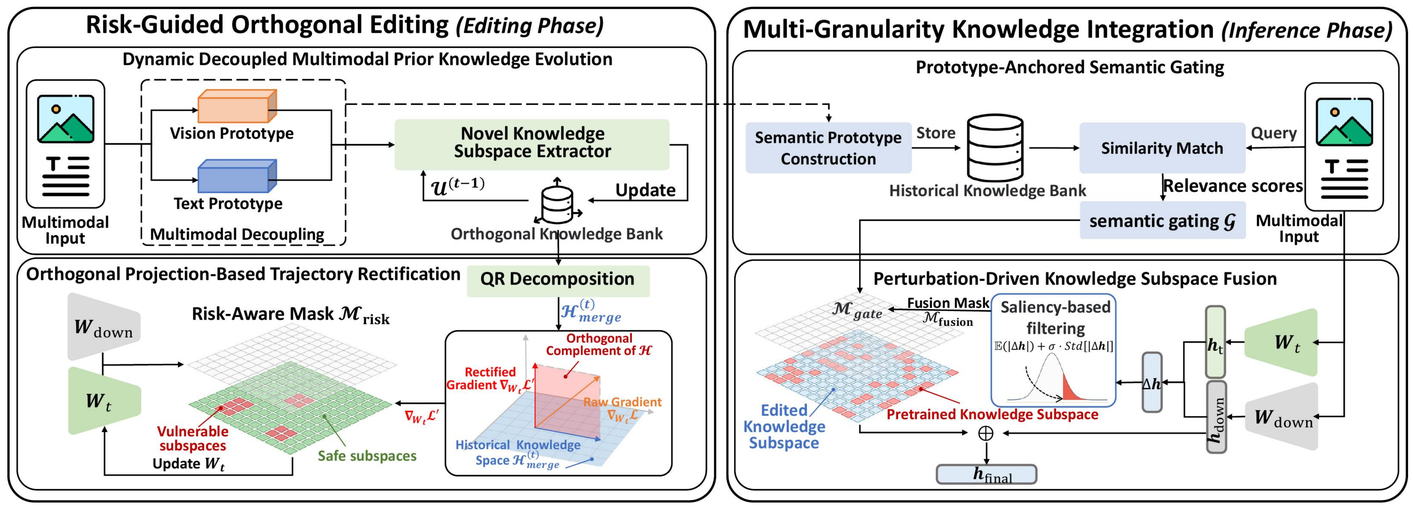


Figure 2. The ROSE editing and inference framework.

Risk-Guided Orthogonal Editing (Editing Phase)

- **Multimodal Decoupling:** Separates visual and textual representations to construct **modality-specific knowledge subspaces**.
- **Orthogonal Projection:** Projects editing gradients onto the **orthogonal complement** of previous knowledge subspaces.
- **Risk-Aware Masking:** Dynamically **constrains parameter updates** to protect pre-trained knowledge.

Multi-Granularity Knowledge Integration (Inference Phase)

- **Prototype-Anchored Gating:** Estimates semantic relevance to **selectively activate the edited branch**.
- **Perturbation-Driven Fusion**: Integrates **salient edited representations** while preserving unrelated pre-trained knowledge.