

CORE CONTRIBUTION

Lifelong MLLM editing without modality entanglement

M-ORE decouples visual/text locality statistics and performs Sherman--Morrison recursive writes in a fixed orthogonal low-rank subspace, enabling stable sequential editing with constant overhead.

DIAGNOSE

Modality-scale mismatch and edit-core reuse reveal distinct sources of online editing degradation.

DECOUPLE

Module-wise preconditioners disentangle visual and textual update statistics.

RECURSE

Sherman--Morrison recursion maintains fixed orthogonal writes with constant-size online memory.

+7.07

Avg. over LiveEdit
LLaVA E-IC, 100 edits

94.64

Reliability
LLaVA E-IC, 100 edits

O(1)

per-edit
stream dependence

512

LoRA rank
LLaVA setting

1 Motivation: online MLLM edits must stay plastic and local

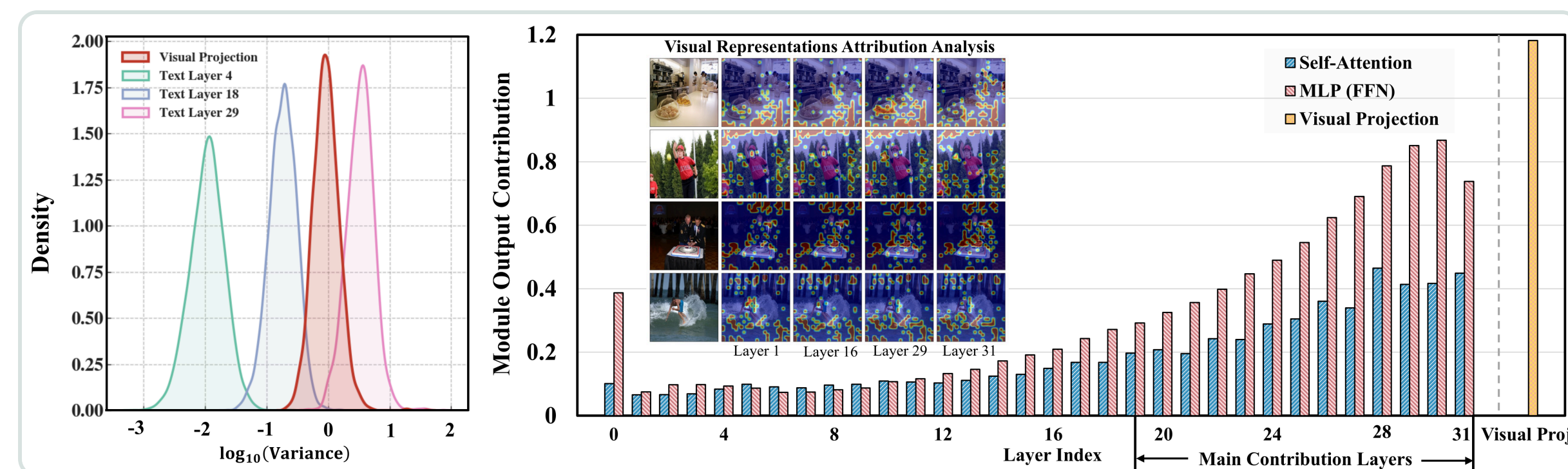
Multimodal LLMs encode visual and textual facts in shared parameters. In deployment, corrections arrive as a stream, but full retraining is infeasible and naive sequential editing accumulates side effects.

- **Reliability**: hit the edited request's target correction.
- **Generality**: transfer to paraphrases and image variants.
- **Locality**: preserve unrelated text and multimodal behavior.

Question. How can MLLMs absorb sequential edits while avoiding modality conflict and edit interference?

2 Failure mode A: cross-modal conflict

Text-oriented editors pool update statistics across heterogeneous MLLM modules. Visual representations carry higher energy, so shared second-order statistics become visually dominated and weaken text-side preservation.



BLIP2-OPT pilot. Visual projection produces higher-scale activations than early text layers and dominates attribution.

17.67

Qwen2-VL
visual output

4.60-11.66

Qwen2-VL
MLP range

51.51

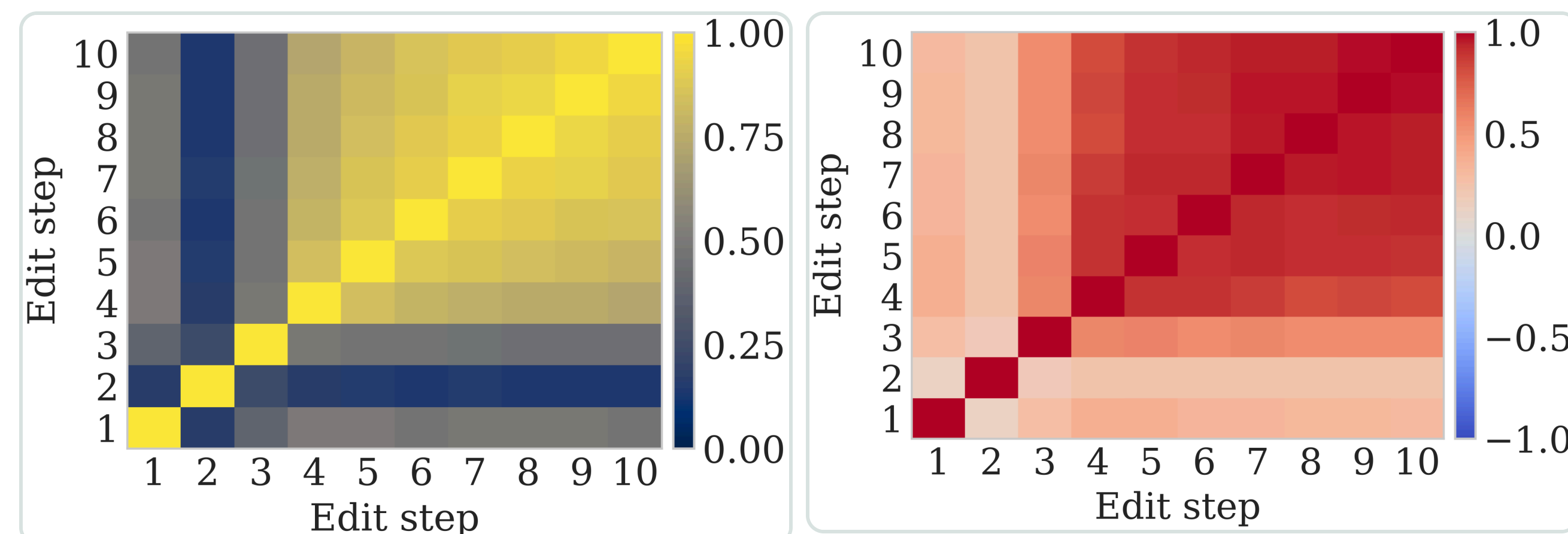
Qwen3-VL
visual output

10.03-25.50

Qwen3-VL
MLP range

3 Failure mode B: inter-edit interference

Sequential parameter-modifying editors repeatedly reuse a compact edit core. Later edits become co-directional with earlier updates, creating long-range coupling and forgetting.



MEND sequential probe. Top-k overlap and cosine similarity show late-stage collapse into co-directional edit cores.

4 M-ORE: decouple statistics, write recursively ★ KEY

DECOUPLE

separate text and visual statistics

FREEZE

orthogonal write basis

UPDATE

Sherman-Morrison recursion

CLOSED-FORM WRITE

$$\Delta B_t^{(l)} = -\eta G_t^{(l)} (I + S_{t-1}^{(l)})^{-1} \triangleq -\eta G_t^{(l)} P_{t-1}^{(l)}. \quad (12)$$

RECURSIVE CACHE

$$P_t^{(l)} = P_{t-1}^{(l)} - \frac{P_{t-1}^{(l)} \xi_t^{(l)} (\xi_t^{(l)})^\top P_{t-1}^{(l)}}{1 + (\xi_t^{(l)})^\top P_{t-1}^{(l)} \xi_t^{(l)}}, \quad P_0^{(l)} = \frac{1}{1+\lambda} I_r. \quad (13)$$

Each edit is preconditioned by past locality statistics, while the current steady feature is recursively folded into the cache for future edits, avoiding edit-growing buffers.

- **Fixed orthogonal coordinates** prevent the write space from drifting with the edit stream.
- **Module-wise preconditioners** keep high-energy visual statistics out of text-side updates.
- **Rank-one sketching** keeps the stored statistic per-edit constant.

5 Main Result: strongest 100-edit averages

+7.07

avg. gain
LLaVA E-IC

94.64

Reliability

93.27

Text
Generality

78.66

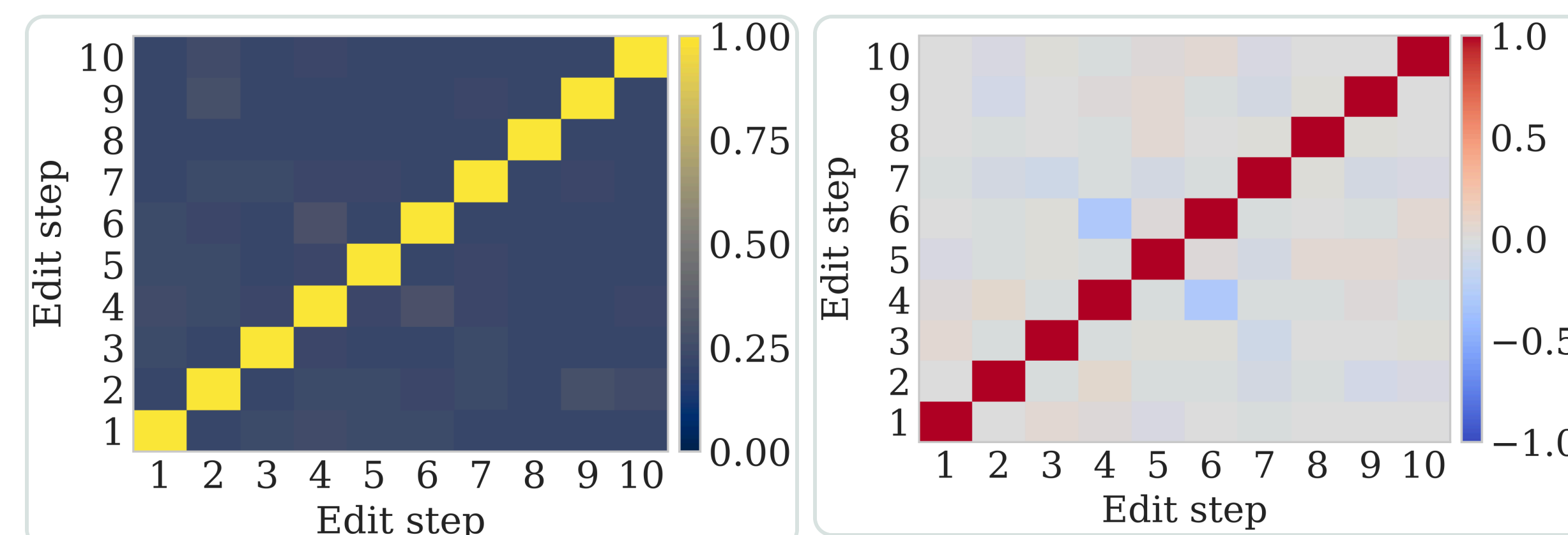
Multimodal
Generality

100-edit setting	LiveEdit	M-ORE	Gain
BLIP2 E-VQA Avg.	92.02	92.56	+0.54
BLIP2 E-IC Avg.	80.60	82.49	+1.89
LLaVA E-VQA Avg.	91.28	93.07	+1.79
LLaVA E-IC Avg.	83.66	90.73	+7.07

LLaVA E-IC. +16.15 Rel., +14.50 T-Gen., +13.16 M-Gen. over LiveEdit.

6 Why it works: no edit-core collapse

Under the same 10-edit interference probe, M-ORE avoids the late-stage reuse pattern observed for MEND.



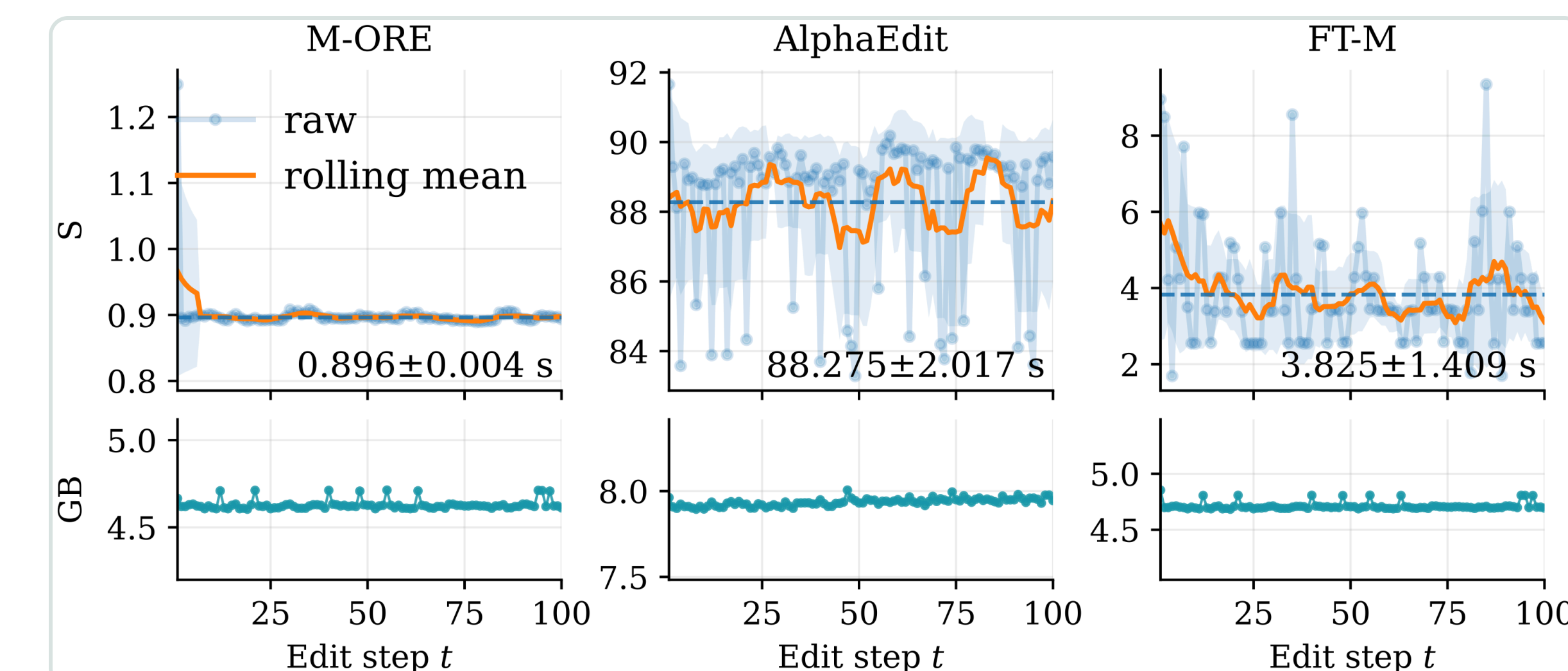
M-ORE sequential probe. Cross-step overlaps stay low and cosine couplings remain near zero.

LLaVA E-IC ablation	Avg.	Δ
M-ORE, 100 edits	90.73	--
w/o frozen A	72.63	-18.10
w/o pooling	92.43	+1.70

Ablation. Frozen A drives stability; pooling is a compact sketch.

7 Efficiency and Scaling

Editor-specific overhead excludes the shared forward/backward pass for edit gradients.



Empirical scaling. After one-time initialization, M-ORE latency and incremental peak memory remain flat over the edit stream.

Editor class	Time/edit	Memory	t-dep.
M-ORE	$\mathcal{O}(\mathcal{L} d_{out} r^2)$	$\mathcal{O}(\mathcal{L} d_{out} r)$	$\mathcal{O}(1)$
Null-space constraint	$\mathcal{O}(d t^2)$	$\mathcal{O}(d t)$	↑
Parameter-preserving	$\mathcal{O}(t)$ or $\mathcal{O}(\log t)$	$\mathcal{O}(t p_{mem})$	↑

8 Generalization to other MLLMs

Using the LLaVA hyperparameters without architecture-specific retuning, M-ORE remains strong on newer MLLMs.

Backbone	#Edit	Rel.	Avg.
Qwen2-VL	1	100.00	97.80
Qwen2-VL	100	93.66	89.50
Qwen3-VL	1	100.00	96.92
Qwen3-VL	100	91.85	89.97

Representation. Locality states stay close through 100/500 edits.

9 Takeaways

DIAGNOSIS. Visual scale mismatch and edit-core reuse explain two failure modes.

METHOD. Frozen orthogonal writes plus module-wise recursive preconditioners.

RESULT. Best 100-edit averages on BLIP2-OPT and LLaVA-v1.5.

COST. Constant stream dependence with compact per-edit state.