Diagnosing and Mitigating Evidence Misgrounding in Multimodal Knowledge Editing

Anonymous ACL submission

Abstract

Multimodal knowledge editing aims to update visually grounded facts in Multimodal Large Language Models (MLLMs) while preserving unrelated text and visual behaviors. However, standard answer-level metrics can hide *evidence misgrounding*: an edit may produce the target answer on the original image yet remain sensitive to incidental context rather than the intended visual evidence. To expose this, we construct WATERBIRD-EDIT, a controlled analysis set that decouples bird habitat labels from background and shows that target-alignment editors can still rely on spurious visual cues. Motivated by this diagnosis, we propose VICE (**V**isual **I**n-**C**ontext **E**diting), a retrieval-based framework that stores edits in external memory and applies them through visual in-context demonstrations that expose task-relevant evidence. Experiments on three standard benchmarks show that VICE achieves strong lifelong editing performance, while results on WATERBIRD-EDIT show that it substantially reduces sensitivity to the measured spurious context.

1 Introduction

Multimodal Large Language Models (MLLMs) have become the backbone of modern vision-language understanding, encoding extensive factual knowledge about the visual world (Li et al., 2023; Zhu et al., 2024; Liu et al., 2023, 2024; Wang et al., 2024b; Bai et al., 2025). When this knowledge becomes outdated or incorrect, multimodal knowledge editing (Cheng et al., 2023; Huang et al., 2024; Chen et al., 2025; Shi et al., 2025) aims to update a visually grounded fact while preserving behavior on unrelated text and visual queries. Unlike text-only editing, a successful MLLM edit must apply the new fact according to the intended visual evidence rather than incidental image context.

However, answer-level success does not guarantee visually grounded editing. In our diagnosis, an editor can produce the corrected answer on the original edit image but become unreliable on image variants that preserve the target evidence while changing incidental context. As illustrated by the bird-habitat example in Figure 1, an edit may succeed on the original water-bird image, yet its prediction can still vary when similar bird evidence appears with a different background. We call this failure mode *evidence misgrounding*: apparent edit success can mask sensitivity to incidental visual context rather than stable alignment with intended evidence. Because standard reliability, generality, and locality metrics mainly measure answer correctness, they can overestimate edit quality by overlooking context-sensitive image generalization.

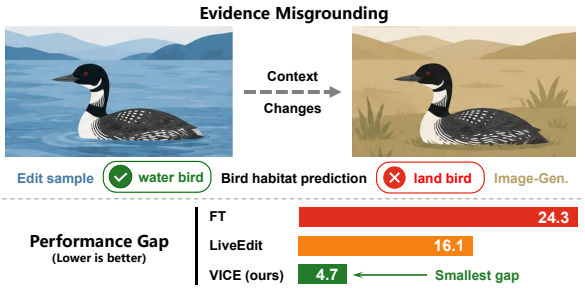


Figure 1: An example of background-induced evidence misgrounding. An edit may succeed on the original image but fail when the same bird evidence appears with a different background. Lower performance gap indicates less sensitivity to background context.

To study this failure in a controlled setting, we use WATERBIRD-EDIT, a diagnostic analysis inspired by Sagawa et al. (2020), where the edited fact is the bird's habitat label and the background provides correlated but potentially spurious context. By decoupling bird habitat and background into four target-context groups, WATERBIRD-EDIT tests whether an edited prediction tracks the bird evidence consistently across background changes. On this diagnosis set, we find that FT and LIVEEDIT (Chen et al., 2025) exhibit behavior consistent with evidence misground-

ing. Although these models increase the likelihood of the desired answer through parameter updates or hidden-state modulation, they do not explicitly separate bird evidence from background context, leaving predictions sensitive to background cues.

These findings suggest that robust MLLM editing should make the evidence used to apply an edit explicit, rather than relying solely on implicit parameter or hidden-state updates. Motivated by this principle, we propose VICE (**V**isual **I**n-**C**ontext **E**diting), a lifelong framework that applies edits through visual in-context learning (Zhou et al., 2024) while freezing the backbone. For each edit, VICE constructs evidence-summary demonstrations whose image summaries highlight task-relevant evidence and separate it from incidental context, then stores the edited fact and its demonstrations in a memory. At inference time, a query retrieves the nearest edit from memory, and a learned scope head decides whether to answer with the retrieved edit context or bypass the edit. This allows the frozen MLLM to apply edits through demonstrations designed to emphasize task-relevant evidence while preserving unrelated behavior.

Our contributions are threefold. First, we identify evidence misgrounding as a failure mode in MLLM editing, and use WATERBIRD-EDIT to diagnose whether edited behavior remains sensitive to incidental context after controlling for the target evidence. Second, we propose VICE, a frozen-backbone framework that applies edits through visual in-context learning. Third, experiments on EVQA (Cheng et al., 2023), EIC (Cheng et al., 2023), VLKEB (Huang et al., 2024), and WATERBIRD-EDIT show that VICE is effective, scalable, and robust to spurious visual context.

2 Diagnosing Evidence Misgrounding

We use evidence misgrounding to denote the case where an edited model yields the corrected answer on the edit image while remaining overly sensitive to incidental visual context rather than consistently following the intended visual evidence. Evidence misgrounding harms edited-model generalization, so detecting this failure can help build robust editing mechanisms. However, existing multimodal editing benchmarks make diagnosis difficult: target objects, scenes, and other visual cues are usually entangled, obscuring whether an edited answer follows the intended evidence or a correlated cue. We therefore construct a controlled diagnostic set with explicit target-context groups to test whether edited behavior follows target evidence across contexts.

2.1 WATERBIRD-EDIT Diagnostic Analysis

To make evidence misgrounding observable, we construct WATERBIRD-EDIT, a controlled diagnostic analysis set built on CUB-200-2011 (Wah et al., 2011) for lifelong editing, inspired by Sagawa et al. (2020).

We first formulate a binary VQA task: given a CUB image, a prompt such as "`What type of bird is in the image? Choose from: land, water.`", the model predicts the species-level bird habitat label $y \in \{\text{water}, \text{land}\}$. Each image is also assigned a background label $b \in \{\text{water}, \text{land}\}$ by a pretrained vision-language model. The habitat and background labels then form four target-context groups $(y, b) \in \{\text{water}, \text{land}\}^2$. In this setting, evidence misgrounding is diagnosed when edited predictions vary with b even though the desired answer is determined by y. To construct the editing samples, we first query the MLLM to be edited on the binary habitat VQA task and identify images where its habitat prediction is incorrect, resulting in 500 such images. Each retained image is then used to form an editing instance, where the target answer is set to the ground-truth habitat label to correct the model's original prediction. The filtered editing samples remain imbalanced across the four target-context groups, thereby preserving the correlation between habitat labels and background labels that the editor may exploit during editing. For the Image-Generality evaluation, each editing sample is paired with another image of the same bird species, which we call image-rephrase sample. This paired image shares the same habitat label as the original editing image but may have a different background, allowing us to evaluate whether the edit generalizes across images of the same species. This design tests whether an editor exposed to spuriously correlated visual cues still generalizes well according to the target bird rather than the background. Details are in Appendix A.

After the 500-edit sequence, we evaluate each editor on the image-rephrase samples and report: (1) I-Gen., the standard Image-Generality metric, (2) Group Image-Generality $\text{IG}(g)$, a fine-grained I-Gen. within each group $g = (y, b)$, (3) Performance Gap (PG), defined as I-Gen. $- \min_g \text{IG}(g)$, and (4) sample-wise normalized probability assigned to the "water" answer among the two candi-

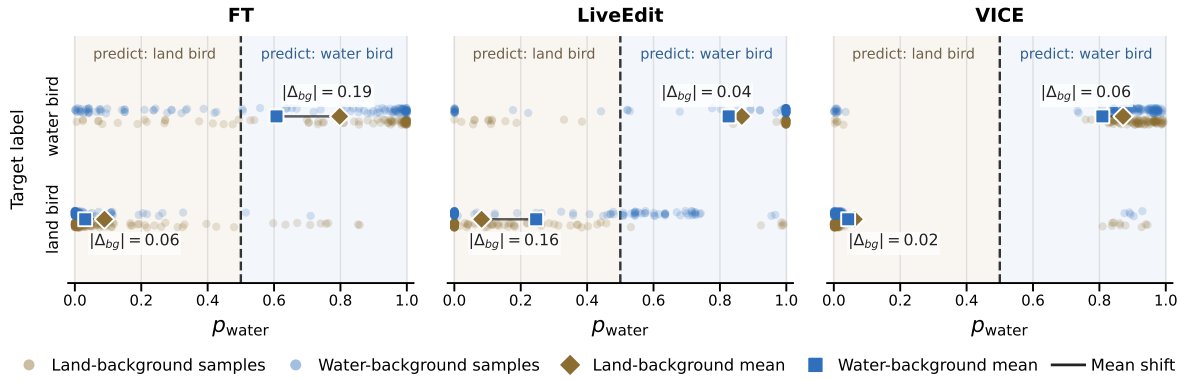


Figure 2: Sample-wise p_{water} defined in Equation (1) for WATERBIRD-EDIT image-rephrase samples after editing. The top two bands show water-bird targets, and the bottom two bands show land-bird targets. Point colors indicate background labels: blue for water backgrounds and orange for land backgrounds. Large markers denote background-group means, and Δ_{bg} reports the background-induced mean shift for each target label.

Method	I-Gen.↑	Mask I-Gen.↑	Group I-Gen.↑				PG↓
			WW	WL	LW	LL	
FT	85.2	87.8	60.9	80.6	98.5	94.3	24.3
LIVEEDIT	83.8	92.4	84.3	84.9	67.7	96.2	16.1
VICE	93.4	89.0	88.7	94.6	96.2	93.7	4.7

Table 1: WATERBIRD-EDIT diagnostics on image-rephrase samples. Each group is written as (habitat, background), where "W" and "L" denote water and land. "PG" denotes performance gap. "Mask I-Gen." is measured on object-only image-rephrase variants that preserve the bird region and mask the background.

date answers ("water" and "land"):

$$p_{\text{water}} = \frac{\exp(s_{\text{water}})}{\exp(s_{\text{water}}) + \exp(s_{\text{land}})}, \qquad (1)$$

where s_{water} and s_{land} are the model logits for the two candidate answers. A large PG indicates target-context sensitivity and functions as a worst-group robustness gap, while the distribution of p_{water} across groups shows how this sensitivity appears at the sample level. We further construct object-only image-rephrase variants using the segmentation masks provided by CUB-200-2011, and report I-Gen. on these variants to measure how edited predictions change after removing the background while preserving the target bird evidence.

2.2 Context Sensitivity After Editing

We focus on two MLLM editors: FT, a direct parameter-update baseline, and LIVEEDIT (Chen et al., 2025), a recent lifelong MLLM editing method that modulates hidden states. The diagnostic experiments lead to the following findings.

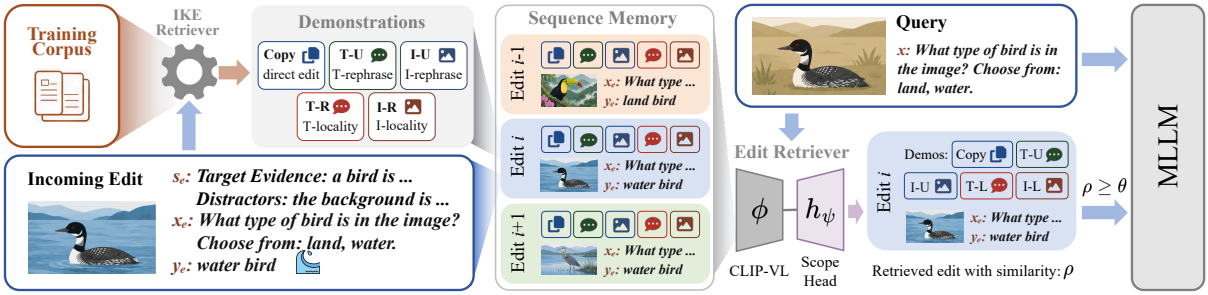
Editors are not robust to target-context shifts. As shown in Table 1, FT and LIVEEDIT achieve reasonable full-set I-Gen. scores of 85.2 and 83.8, yet their worst groups are much lower, producing large PG values of 24.3 and 16.1. This gap shows that their apparent image generalization is not robust under changes in the incidental background context. Figure 2 shows the same pattern at the probability level: for a fixed target label, changing the background substantially shifts p_{water}, especially for FT on water targets ($|\Delta_{bg}| = 0.19$) and LIVEEDIT on land targets ($|\Delta_{bg}| = 0.16$).

Object-only masking disentangles the effect of background context. Object-only masking provides a complementary diagnostic by preserving the target bird region while removing the surrounding background. Table 1 shows that masking improves FT from 85.2 to 87.8 and LIVEEDIT from 83.8 to 92.4, indicating that background context harms their original image-generality performance.

Together, these diagnostics suggest that robust edit application should make the relevant visual evidence explicit, rather than relying solely on internal parameter or hidden-state changes. This motivates a visual in-context editing framework that presents retrieved edit demonstrations at inference time, allowing the frozen MLLM to apply the edit from explicit visual evidence.

3 Method: VICE

Building on the above diagnosis, we propose VICE, a frozen-backbone framework that applies edits through retrieved visual demonstrations at inference time. Each edit is stored with



(a) Evidence-Summary Demonstrations (b) Lifelong Editing (c) Inference

Figure 3: Illustration of the VICE pipeline. (a) Evidence-summary demonstrations for Copy, Text-Update (T-U), Image-Update (I-U), Text-Retain (T-R), and Image-Retain (I-R), all using the same template in Equation (3). (b) Lifelong editing, where each incoming new fact is stored together with its summary and retrieved demonstrations in Sequence Memory. (c) Inference, where a multimodal encoder retrieves the most relevant edit and routes the query augmented by the retrieved context to the MLLM when the retrieval similarity score exceeds the threshold.

evidence-summary demonstrations that highlight task-relevant visual cues and separate them from incidental context. Given a query, VICE retrieves the most relevant edit and uses scope-aware routing to decide whether the edit should be applied. Figure 3 illustrates the VICE pipeline.

3.1 Preliminary: In-Context Editing

Zheng et al. (2023) propose IKE and frame editing as prompting with demonstrations rather than updating parameters. Given a new fact (x_e, y_e) and a query x, IKE forms a test-time prompt as "$\mathcal{C}$. New Fact: x_e y_e. Prompt: x", where $\mathcal{C}$ contains retrieved demonstrations that teach when to copy, update, or retain answers. Each demonstration $c_k \in \mathcal{C}$ uses the same two textual fields:

$$c_k = \text{“New Fact: } x_e^k \; y_e^k\text{. Prompt: } \tilde{x}^k \; \tilde{y}^k\text{”}, \quad (2)$$

where both the demonstration fact (x_e^k, y_e^k) and the prompt-answer pair $(\tilde{x}^k, \tilde{y}^k)$ are drawn from training-set examples. IKE retrieves these demonstrations using sentence-level similarity between the requested new fact and candidate training entries. We follow this retrieval-based protocol for selecting demonstration content. For MLLM editing, VICE keeps the "New Fact / Prompt" template but adds an "Image Summary" field, introduced next, so that demonstrations can expose textualized visual evidence for the demonstrated fact.

3.2 Evidence-Summary Demonstrations

Motivated by a recent visual in-context learning method (Zhou et al., 2024), VICE extends IKE's textual demonstrations by adding an evidence-summary field for multimodal edits. The goal is to make the visual evidence associated with the demonstrated fact explicit in the prompt, rather than leaving the "New Fact" as an answer-only association. Starting from IKE's "New Fact / Prompt" demonstration, VICE inserts an "Image Summary" field before the "New Fact":

Type	Image Summary	New Fact	Prompt
Copy	$s_e = G(i_e, x_e)$	(x_e, y_e)	(x_e, y_e)
Text-Update	$s_e = G(i_e, x_e)$	(x_e, y_e)	(x_r, y_e)
Image-Update	$s_r = G(i_r, x_e)$	(x_e, y_e)	(x_e, y_e)
Text-Retain	$s_e = G(i_e, x_e)$	(x_e, y_e)	(x_l, y_l)
Image-Retain	$s_l = G(i_l, x_{il})$	(x_e, y_e)	(x_{il}, y_{il})

Table 2: Five evidence-summary demonstration roles instantiated from the template in Equation (3). (i_e, x_e, y_e) is the anchor example; x_r is a text rephrase with answer y_e; i_r is an image variant for prompt x_e; (x_l, y_l) is a text-locality example; and (i_l, x_{il}, y_{il}) is an image-locality example. Superscript of the k-th demonstration c_k is omitted for readability.

$$c_k = \text{“Image Summary: } s^k\text{.} \\ \text{New Fact: } x_e^k \; y_e^k\text{. Prompt: } \tilde{x}^k \; \tilde{y}^k\text{”}. \quad (3)$$

This keeps the textual interface of in-context editing while allowing a demonstration to state the image evidence associated with its own "New Fact" field. Concretely, given an edit image and prompt (i_e^k, x_e^k), we use a frozen summary model G to generate a text-guided image summary

$$s_e^k = G(i_e^k, x_e^k), \quad (4)$$

where s_e^k describes object evidence relevant to x_e^k and separates it from incidental visual attributes. The generation prompt explicitly instructs the summary model not to answer the question directly,

so the summary remains evidence-focused rather than revealing or restating the edit target. In our implementation, G can be instantiated as the same pretrained MLLM before editing, which avoids introducing a stronger external teacher. Please refer to Appendix B for generation details.

Table 2 summarizes five demonstration roles. "Copy" shows direct use of the "New Fact". "Text-Update" and "Image-Update" show transfer to a rephrased question or a related image, respectively. "Text-Retain" and "Image-Retain" pair the "New Fact" with unrelated queries whose original answers should be preserved.

3.3 Lifelong Visual In-Context Editing

In the lifelong setting, edits arrive sequentially. For each edit (i_e, x_e, y_e), VICE builds the image summary $s_e = G(i_e, x_e)$, retrieves the evidence-summary demonstrations $\mathcal{C}$ described in Section 3.2, following IKE's sentence-level similarity retrieval protocol, and appends the resulting record to an external **Sequence Memory** $\mathcal{M}$:

$$\mathcal{M} \leftarrow \mathcal{M} \cup \{(i_e, x_e, y_e, s_e, \mathcal{C})\}. \tag{5}$$

Editing is therefore a simple memory-write operation and does not modify model parameters or hidden states.

At inference time, VICE decides whether a query should use a stored edit. Retrieval first proposes a memory entry, and scope verification then decides whether to apply it. We construct a retrieval encoder from a frozen CLIP-based multimodal encoder ϕ and a learnable scope-discriminative head h_ψ. The encoder ϕ concatenates image and text features, while h_ψ projects them into a new space that separates in-scope and out-of-scope knowledge. Given a query (i, x), VICE retrieves the most similar memory entry:

$$m^* = \arg\max_{m \in \mathcal{M}} \mathbf{z}^\top \mathbf{z}_m, \tag{6}$$

where $\mathbf{z} = h_\psi(\phi(i, x)), \mathbf{z}_m = h_\psi(\phi(i_m, x_m))$, i_m, x_m are entries in a cached memory m, and $m^* = (i_e^*, x_e^*, y_e^*, s_e^*, \mathcal{C}^*)$. We use $\rho = \max_{m \in \mathcal{M}} \mathbf{z}^\top \mathbf{z}_m$ as the retrieval score. Given a fixed threshold θ, if $\rho \geq \theta$, the retrieved edit is applied to form an augmented prompt:

$$\begin{aligned} &\text{“}\mathcal{C}^*\text{. Image Summary: } s_e^*. \\ &\qquad \text{New Fact: } x_e^*\ y_e^*\text{. Prompt: } x\text{”}, \end{aligned} \tag{7}$$

along with the original image input i to query the MLLM. Otherwise, the model receives only the original query without any retrieved edit information. Algorithm 1 summarizes the inference.

Algorithm 1 VICE Inference

Require: Frozen MLLM, sequence memory $\mathcal{M}$, multimodal encoder ϕ, scope-discriminative head h_ψ, threshold θ, query (i, x)
Ensure: Prediction $\hat{y}$
  $\mathbf{z} \leftarrow h_\psi(\phi(i, x))$
  $(i_e^*, x_e^*, y_e^*, s_e^*, \mathcal{C}^*), \rho \leftarrow$ Equation (6)
  if $\rho \geq \theta$ **then**
    $p \leftarrow$ Equation (7)
    $\hat{y} \leftarrow \text{MLLM}(i, p)$
  else
    $\hat{y} \leftarrow \text{MLLM}(i, x)$
  end if
  return $\hat{y}$

Learnable scope-discriminative head. The scope-discriminative head h_ψ is a two-layer MLP and is trained offline, separately from the editing and inference pipeline, while both CLIP and the MLLM remain frozen. Training uses no test edit anchors or evaluation queries. For each training edit group, the reliability entry is used as the anchor, text/image-rephrase samples are treated as in-scope examples, and text/image-locality samples are treated as out-of-scope examples. Given minibatch anchors $\mathcal{A}$, non-anchor in-scope examples $\mathcal{R}$, and locality examples $\mathcal{O}$, we optimize an anchor retrieval loss:

$$\mathcal{L}_{\text{ret}} = \frac{1}{|\mathcal{R}|} \sum_{q \in \mathcal{R}} -\log \frac{\exp(\mathbf{z}_q^\top \mathbf{a}^*(q)/T)}{\sum_{j \in \mathcal{A}} \exp(\mathbf{z}_q^\top \mathbf{a}_j/T)}, \tag{8}$$

where $\mathbf{a}^*(q)$ is the ground-truth anchor for q and T is the contrastive temperature. We also use a margin loss to push locality examples away from all anchors:

$$\mathcal{L}_{\text{margin}} = \frac{1}{|\mathcal{O}|} \sum_{q \in \mathcal{O}} \max\big(0, \max_{j \in \mathcal{A}} \mathbf{z}_q^\top \mathbf{a}_j - m\big). \tag{9}$$

The total objective is $\mathcal{L}_{\text{scope}} = \mathcal{L}_{\text{ret}} + \mathcal{L}_{\text{margin}}$. After training, the dataset-specific h_ψ is fixed and used only for retrieval and threshold routing. Training details are provided in Appendix B.

4 Experiments

We evaluate VICE across three multimodal knowledge editing benchmarks and two vision-language backbones, comparing against five representative editing baselines. Section 4.1 describes the experimental setup, including datasets, backbones,

Model	# Edits	Method	EVQA					EIC					VLKEB				
			Rel	T-Gen	I-Gen	T-Loc	I-Loc	Rel	T-Gen	I-Gen	T-Loc	I-Loc	Rel	T-Gen	I-Gen	T-Loc	I-Loc
LLaVA-v1.5	1	FT-VL	100.00	91.03	95.38	95.29	90.53	**99.93**	**94.65**	**97.34**	91.63	82.16	**100.00**	**98.74**	**99.92**	76.13	84.15
		TP	33.27	31.33	29.06	96.77	98.83	57.49	57.48	53.91	95.41	96.52	50.34	52.43	50.27	92.59	96.58
		MEND	91.23	90.05	91.29	91.02	90.22	92.82	91.81	90.59	96.38	93.69	92.13	91.28	90.22	89.19	90.13
		SERAC	89.33	83.72	84.97	82.05	23.78	88.18	81.03	85.61	84.01	28.58	89.77	89.11	87.92	66.68	14.20
		LiveEdit	94.28	**94.51**	88.01	100.00	100.00	82.16	81.01	78.27	100.00	100.00	96.43	95.22	93.72	100.00	**100.00**
		VICE	**100.00**	87.33	**97.57**	**100.00**	**100.00**	98.91	92.76	79.61	**100.00**	**100.00**	99.98	98.27	96.72	**100.00**	99.98
	100	FT-VL	76.44	67.30	67.76	58.04	46.18	73.11	69.78	65.46	52.21	58.69	75.91	72.95	74.89	56.36	70.29
		TP	32.60	30.74	28.65	83.13	89.68	50.31	50.67	46.83	34.35	63.78	48.26	50.78	48.20	64.18	79.15
		MEND	2.22	2.20	2.21	0.21	0.62	56.83	56.89	53.07	87.63	84.64	0.56	0.58	0.66	0.18	0.07
		SERAC	88.08	81.53	82.48	62.13	12.90	53.35	53.70	49.41	48.04	17.25	72.25	62.43	70.68	53.73	13.69
		LiveEdit	93.54	**92.34**	85.89	100.00	99.31	80.81	78.77	63.52	100.00	**100.00**	94.56	90.65	89.56	100.00	**100.00**
		VICE	**100.00**	87.15	**97.50**	**100.00**	**99.82**	**98.91**	**92.76**	**73.97**	**100.00**	99.95	**99.98**	**98.27**	**97.14**	**100.00**	99.85
	1000	FT-VL	60.21	50.76	51.03	24.37	29.48	64.21	59.43	57.09	33.39	38.30	65.30	62.67	63.87	33.89	37.32
		TP	32.82	30.48	28.55	42.12	60.28	33.75	36.27	32.34	10.47	26.22	34.23	34.74	34.44	27.11	45.52
		MEND	0.04	0.05	0.05	0.08	0.09	54.39	54.14	50.99	83.87	80.60	0.03	0.05	0.07	0.06	0.08
		SERAC	85.57	75.58	82.01	62.46	15.69	52.93	53.44	49.01	49.91	16.65	60.93	56.49	60.06	52.94	15.04
		LiveEdit	92.93	**90.16**	84.30	100.00	96.43	72.80	69.95	57.05	100.00	99.79	92.22	83.97	82.75	100.00	**100.00**
		VICE	**100.00**	87.03	**97.01**	**100.00**	**98.38**	**98.91**	**92.76**	**60.91**	**100.00**	**99.90**	**99.98**	**98.26**	**96.75**	**100.00**	98.32
MiniGPT4	1	FT-VL	**100.00**	**99.74**	**99.86**	63.66	35.65	**100.00**	**95.28**	**96.41**	58.21	33.84	**100.00**	**99.75**	**99.95**	56.12	42.03
		TP	30.07	28.55	30.40	89.48	71.50	30.79	32.00	34.41	44.72	28.74	37.17	38.31	37.49	74.56	46.15
		MEND	93.47	91.32	91.22	81.51	74.97	90.28	83.91	79.05	95.92	93.97	90.28	89.18	89.55	93.63	86.69
		SERAC	87.66	69.24	85.96	82.11	24.11	85.09	78.96	81.70	83.36	24.50	89.28	89.06	87.35	64.88	12.72
		LiveEdit	94.90	92.63	92.06	100.00	100.00	86.84	83.82	76.95	100.00	100.00	96.41	95.56	92.85	100.00	**100.00**
		VICE	99.98	88.74	97.35	**100.00**	**100.00**	98.93	93.67	77.85	**100.00**	**100.00**	95.27	82.95	92.05	**100.00**	99.99
	100	FT-VL	56.02	49.77	50.35	16.97	10.42	59.24	56.32	52.58	4.13	4.65	64.49	63.10	62.70	21.95	14.10
		TP	6.64	6.16	6.04	1.80	0.61	7.74	7.44	7.23	1.08	0.63	4.93	4.85	4.40	1.57	1.34
		MEND	23.32	20.75	21.42	47.27	42.60	51.92	53.96	48.24	86.02	83.88	38.85	40.02	37.99	72.82	68.89
		SERAC	87.39	66.53	81.70	65.13	14.04	47.41	46.17	45.22	49.37	13.52	69.38	58.31	67.57	51.71	13.72
		LiveEdit	93.23	**91.27**	88.34	100.00	99.61	84.38	83.53	**71.85**	100.00	**100.00**	94.06	**94.62**	89.42	100.00	**100.00**
		VICE	**99.98**	88.49	**97.28**	**100.00**	**99.82**	**98.93**	**93.67**	71.73	**100.00**	99.95	**95.27**	82.95	**92.44**	**100.00**	99.87
	1000	FT-VL	36.32	31.16	34.89	5.27	5.86	62.25	57.95	54.73	7.16	11.06	42.64	41.62	41.18	8.55	9.32
		TP	7.70	6.99	7.50	1.70	0.55	8.16	8.26	7.63	0.59	1.46	3.74	3.47	3.57	1.17	2.75
		MEND	31.84	26.60	33.98	43.08	44.90	49.30	51.46	46.70	81.03	82.53	42.35	45.07	42.68	66.98	62.32
		SERAC	84.50	60.36	81.83	63.26	13.40	47.03	45.79	44.82	42.12	15.31	57.02	51.66	56.80	48.48	13.24
		LiveEdit	92.75	**89.81**	85.12	100.00	96.82	81.43	80.34	**62.25**	100.00	99.73	93.41	**89.87**	85.21	100.00	**100.00**
		VICE	**99.98**	88.37	**96.82**	**100.00**	**98.26**	**98.93**	**93.67**	57.31	**100.00**	**99.92**	**95.27**	82.95	**92.14**	**100.00**	98.29

Table 3: Main results of lifelong editing for LLaVA-v1.5 and MiniGPT4 on EVQA, EIC, and VLKEB. "#Edits" represents the number of new facts. MEND, SERAC, and LiveEdit results are from Chen et al. (2025); FT-VL, TP, and VICE are from our runs.

baselines, evaluation metrics, and implementation details. Section 4.2 presents the main evaluation results, and Section 4.4 reports our ablation studies.

4.1 Experimental Setup

Datasets and Backbones. We evaluate on three standard multimodal knowledge editing benchmarks: **EVQA** (Cheng et al., 2023), **EIC** (Cheng et al., 2023), and **VLKEB** (Huang et al., 2024). We use two MLLMs as the backbone: **LLaVA-v1.5-7B** (Liu et al., 2024) and **MiniGPT-4-vicuna-7B** (Zhu et al., 2024). Both are widely used in the multimodal knowledge editing literature and represent different architectural choices.

Baselines. We compare VICE against five representative knowledge editing methods: **FT-VL**, fine-tuning the final layer of the MLLM; **TP** (Huang et al., 2023), appending transformer-patchers to FFNs; **MEND** (Mitchell et al., 2022a), training a hyper-network to learn to update model parameters; **SERAC** (Mitchell et al., 2022b), a representative memory-based method; **LiveEdit** (Chen et al., 2025), a recent multimodal editing baseline that modulates hidden-states with MoE modules.

Metrics. Following Cheng et al. (2023), we report five standard metrics: **Reliability**: whether the model produces the target answer on the edited input; **Text/Image-Locality**: whether unrelated text and multimodal queries remain unchanged; **Text/Image-Generality**: whether the edit transfers to textually rephrased and visually similar variants.

Implementation details. For VICE, all MLLM backbone parameters are frozen throughout editing and inference. We use CLIP-ViT-B/32 as ϕ and train only the two-layer head h_ψ with 256-

# Edits	Ablation	Rel	T-Gen	I-Gen	T-Loc	I-Loc
1	Full VICE	**1.00**	**0.88**	**0.97**	**1.00**	**0.99**
	w/o Scope Head	1.00	0.86	0.96	1.00	0.85
	w/o Image Summary	0.75	0.46	0.80	1.00	1.00
	w/o Retrieved Demos	0.32	0.30	0.27	1.00	1.00
1000	Full VICE	**1.00**	**0.88**	**0.97**	**1.00**	**0.98**
	w/o Scope Head	1.00	0.85	0.95	1.00	0.76
	w/o Image Summary	0.75	0.46	0.80	0.99	1.00
	w/o Retrieved Demos	0.32	0.30	0.27	1.00	1.00

Table 4: Ablation study of VICE on EVQA with LLaVA-v1.5. "w/o Scope Head" uses the raw retrieval route without the learned head h_ψ, "w/o Image Summary" removes the generated visual evidence summaries from demonstrations, and "w/o Retrieved Demos" removes retrieved visual in-context demonstrations.

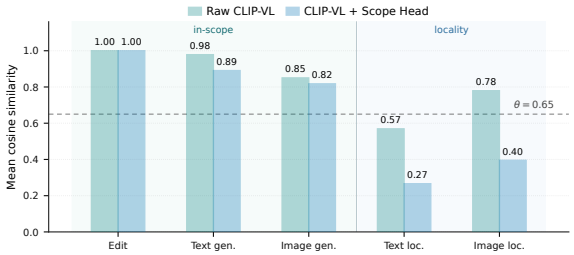


Figure 4: Illustration of mean retrieval similarity of raw CLIP-VL features and learned scope-head embeddings on EVQA. For edit, text-rephrase, and image-rephrase queries, we report similarity to the paired edit anchor; for locality, we report nearest-anchor similarity.

dimensional outputs. The head is trained offline for each dataset with AdamW for 80 epochs, batch size 32, learning rate 10^{-3}, temperature $T = 0.07$, and margin $m = 0.5$. Image summaries are generated before editing and cached. We use the fixed threshold $\theta = 0.65$ for all main experiments.

4.2 Main Results

As demonstrated in Table 3, VICE achieves strong and stable performance across diverse lifelong editing setups. Its main gains appear in reliability, image-generality, and locality, which are central to visually grounded editing. For example, under 1000 edits for LLaVA-v1.5, VICE improves I-Gen. over LIVEEDIT by 12.7 on EVQA, 3.9 on EIC, and 14.0 on VLKEB, and for MINIGPT4, by 11.7 on EVQA and 6.9 on VLKEB. VICE is not uniformly best on text-generality, where LIVEEDIT remains competitive in some settings, but it maintains strong T-Gen., while giving more consistent image-side generalization and locality. More importantly, VICE remains highly stable as the number of edits increases, showing little performance degradation across both backbones and benchmarks. By contrast, the baselines reveal a clear scalability–locality trade-off. FT-VL often fits isolated edits but quickly harms locality under long sequences, while TP has limited capacity for complex multimodal knowledge. MEND and SERAC are stronger in some single-edit settings, but MEND's accumulated updates scale poorly and SERAC's text-based retrieval gives weak image locality. LiveEdit is the strongest recent baseline in multimodal knowledge editing, yet it still declines as edits accumulate, whereas VICE remains more stable and therefore better suited for lifelong multimodal knowledge editing.

4.3 Robustness on WATERBIRD-EDIT

The diagnosis in Section 2 shows that FT and LIVEEDIT can achieve reasonable average I-Gen. while still varying strongly across target-context groups. Applying the same analysis to VICE, Table 1 shows a different pattern: VICE reaches the highest overall I-Gen. (93.4) and the smallest PG (4.7). Its group accuracies also remain consistently high, rather than showing a sharp worst-group drop. This suggests that VICE improves not only average I-Gen., but also worst-group robustness, reducing sensitivity to background context.

Object-only masking provides a complementary check. When the background is removed and the bird region is preserved, VICE remains strong (89.0) and does not show the large masking benefit observed for the baselines. This indicates that its original image-rephrase performance is less degraded by spurious context. Together, the group-level and masking results support the claim that visual in-context editing reduces evidence misgrounding on WATERBIRD-EDIT.

4.4 Ablation Study

To assess the contribution of each component, we ablate the scope head, image summaries, and retrieved demonstrations on EVQA with LLaVA-v1.5. As shown in Table 4, removing the scope head mainly reduces I-Loc., which suggests that scope routing helps prevent irrelevant edit application. Removing image summaries lowers Rel., T-Gen., and I-Gen., showing that explicit visual evidence is important for generalizing the edit to rephrased or visually similar inputs. The largest drop occurs when retrieved demonstrations are removed, indicating that they provide the main in-context editing signal. Overall, the ablation results

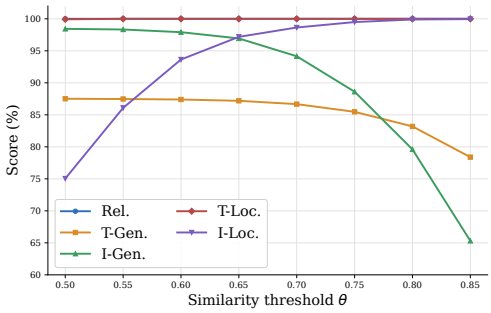


Figure 5: Five standard metrics under different similarity thresholds θ on EVQA with LLaVA-v1.5.

align with the intended roles of the three components: retrieved demonstrations convey the edit, image summaries ground it in visual evidence, and the scope-discriminative head preserves locality.

Figure 4 further supports this interpretation: the learned scope head creates a clearer margin between in-scope and out-of-scope facts, whereas raw CLIP-VL assigns some locality queries similarities close to in-scope examples, making them more likely to trigger irrelevant edits.

4.5 Sensitivity to Similarity Threshold

We evaluate the sensitivity of VICE to the similarity threshold θ by sweeping θ from 0.50 to 0.85 on EVQA with LLaVA-v1.5. As shown in Figure 5, reliability and text-locality are stable across the sweep, while image-generality and image-locality show the main trade-off. A low threshold applies edits too broadly, giving high I-Gen. (98.4 at $\theta = 0.50$) but low I-Loc. (75.0). A very high threshold is too conservative, causing I-Gen. to drop to 65.3 at $\theta = 0.85$. We use $\theta = 0.65$ in the main experiments because it gives a good balance, with I-Gen. 96.9 and I-Loc. 97.2. We also keep the same $\theta = 0.65$ for other main experiments to avoid additional hyperparameter tuning.

5 Related Work

Knowledge Editing. Knowledge editing aims to update specific factual associations in pretrained models while preserving their behavior on unrelated inputs. Prior work edits models by rewriting located parameters (Dai et al., 2022; Meng et al., 2022a,b; Fang et al., 2025), learning parameter updates (Sinitsin et al., 2020; De Cao et al., 2021; Mitchell et al., 2022a; Tan et al., 2024), adding trainable modules (Huang et al., 2023; Wang and Li, 2024), or using external memory and routing (Mitchell et al., 2022b; Hartvigsen et al., 2023; Yu et al., 2024; Wang et al., 2024c; Zeng et al., 2025; Xia et al., 2025). In-context editing (Si et al., 2022; Madaan et al., 2022; Zheng et al., 2023; Ma et al., 2025) instead specifies new facts through demonstrations at inference time, avoiding backbone updates. To achieve lifelong editing (Chen et al., 2025, 2024; Pan et al., 2024), various recent works have incorporated lifelong learning techniques (Wang et al., 2024a).

Multimodal knowledge editing (Chen et al., 2025; Shi et al., 2025) extends these ideas to vision-language models. MMEdit (Cheng et al., 2023) and VLKEB (Huang et al., 2024) provide benchmarks. These evaluations mainly test answer correctness, not whether the edited answer is grounded in the intended visual evidence. Our method VICE addresses this gap with visual in-context demonstrations and a lightweight scope router.

In-Context Learning. In-context learning (Brown et al., 2020; Liu et al., 2022; Bu et al., 2024, 2025) enables large language models to adapt to new tasks from demonstrations without parameter updates. Vision-language models can also learn from interleaved image-text demonstrations, though cross-modal interaction makes visual in-context learning more challenging (Schwettmann et al., 2023). Recent VICL work retrieves visual demonstrations and summarizes images according to task intent (Zhou et al., 2024). We adapt this idea to multimodal knowledge editing: VICE uses visual demonstrations to express edited facts and a scope head to decide when they should apply.

6 Conclusion

We studied multimodal knowledge editing beyond answer-level success, asking whether edited predictions follow the intended visual evidence. Standard metrics can hide evidence misgrounding: a model may produce the target answer while remaining sensitive to incidental context. To make this visible, we introduced WATERBIRD-EDIT, which separates bird-habitat facts from background cues, and VICE, a frozen-backbone framework that stores edits in sequence memory and applies them through visual in-context demonstrations with scope routing. Across lifelong editing benchmarks, VICE remains strong; on WATERBIRD-EDIT, it substantially reduces sensitivity to spurious background context. These results show that explicitly exposing target visual evidence helps make multimodal edits more grounded and scalable.

Limitations

WATERBIRD-EDIT is designed as a controlled diagnostic setting for bird-habitat editing, where target evidence and background context can be explicitly separated. This makes the evidence-misgrounding pattern measurable, while future work can extend the same diagnostic principle to broader visual facts and more open-ended scenes. VICE currently uses generated image summaries and a learned retrieval-based scope decision to expose and apply edits. These components are effective in our experiments, and further improvements in summary quality, retrieval calibration, and uncertainty handling may make the framework more reliable in ambiguous boundary cases. Finally, VICE shifts part of the editing burden from backbone updates to external memory, retrieval, and longer prompts. This trade-off preserves the frozen MLLM and supports lifelong editing, while leaving useful opportunities for memory compression, faster indexing, and more compact prompt construction.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Taiji Suzuki, Qingfu Zhang, and Hau-San Wong. 2024. Provably transformers harness multi-concept word semantics for efficient in-context learning. *Advances in Neural Information Processing Systems*, 37:63342–63405.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Qingfu Zhang, Hau-San Wong, and Taiji Suzuki. 2025. Provable in-context vector arithmetic via retrieving task concepts. *arXiv preprint arXiv:2508.09820*.

Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. 2025. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9455–9466.

Qizhou Chen, Taolin Zhang, Xiaofeng He, Dongyang Li, Chengyu Wang, Longtao Huang, and 1 others. 2024. Lifelong knowledge editing for llms with retrieval-augmented continuous prompt learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13565–13580.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023. Can we edit multimodal large language models? In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 13877–13888.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6491–6506.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, volume 2025, pages 16366–16396.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36:47934–47959.

Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Vlkeb: A large vision-language model knowledge editing benchmark. *Advances in Neural Information Processing Systems*, 37:9257–9280.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. Transformer-patcher: One mistake worth one neuron. In *International Conference on Learning Representations*, pages 13927–13943.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pages 100–114.

Yaohui Ma, Xiaopeng Hong, Shizhou Zhang, Huiyun Li, Zhilin Zhu, Wei Luo, and Zhiheng Ma. 2025. Comprehendedit: a comprehensive dataset and evaluation framework for multimodal knowledge editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19323–19331.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2833–2861.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*, pages 11576–11597.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831.

Kaihang Pan, Zhaoyu Fan, Juncheng Li, Qifan Yu, Hao Fei, Siliang Tang, Richang Hong, Hanwang Zhang, and Qianru Sun. 2024. Towards unified multimodal editing with enhanced knowledge collaboration. *Advances in Neural Information Processing Systems*, 37:110290–110314.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*, pages 2661–2680.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. 2023. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2862–2867.

Zhiyi Shi, Binjie Wang, Chongjie Si, Yichen Wu, Junsik Kim, and Hanspeter Pfister. 2025. Dualedit: Dual editing for knowledge updating in vision-language models. *arXiv preprint arXiv:2506.13638*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, and Artem Babenko. 2020. Editable neural networks. *arXiv preprint arXiv:2004.00345*.

Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language models via meta learning. In *International Conference on Learning Representations*, volume 2024, pages 54676–54690.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The Caltech-Ucsd Birds-200-2011 Dataset.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024a. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2024c. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797.

Renzhi Wang and Piji Li. 2024. Lemoe: Advanced mixture of experts adaptor for lifelong model editing of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2551–2575.

Shujun Xia, Haokun Lin, Yichen Wu, Yinan Zhou, Zixuan Li, Zhongwei Wan, Xingrun Xing, Yefeng Zheng, Xiang Li, Caifeng Shan, and 1 others. 2025. Medrek: Retrieval-based editing for medical llms with key-aware prompts. *arXiv preprint arXiv:2510.13500*.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. 2024. Melo: Enhancing model editing with neuron-indexed dynamic lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19449–19457.

Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. 2025. Visual-oriented fine-grained knowledge editing for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2491–2500.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. 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.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15890–15902.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, and 1 others. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

A WATERBIRD-EDIT Construction Details

Source data and labels. WATERBIRD-EDIT is constructed from CUB-200-2011, which provides fine-grained bird images with species identities. We convert the original species labels into a binary habitat label $y \in \{\text{water}, \text{land}\}$. We obtain the background label $b \in \{\text{water}, \text{land}\}$ by using a pre-trained vision-language model to judge whether the surrounding scene contains visible water. Specifically, we query LLaVA-v1.6-Vicuna-13B on each CUB image with a binary prompt asking whether the background contains any water body, such as a river, lake, ocean, pond, waterfall, or stream. The model is instructed to answer only "Yes" or "No"; we map "Yes" to a water background and "No" to a land background. The generated file contains one background label for each of the 11,788 CUB images. These background labels are used only for dataset construction and group-wise diagnostics, and are not included in the edit prompt. This gives an explicit target-context grouping (y, b), where the first component is the bird habitat label and the second component is the surrounding background. The construction is designed so that the edited answer should follow the bird identity rather than the background scene.

Background-labeling prompt. The key prompt used for assigning background labels is:

```
Please carefully examine the background
of this image and determine whether
there is any water present (such
as rivers, lakes, oceans, ponds,
waterfalls, streams, or any other form
of water). Please respond with only
"Yes" or "No", without any additional
explanations or descriptions. If you
can see any water body in the background,
answer "Yes"; if there is no water body
visible in the background, answer "No".
```

Edit-image filtering. We start from a filtered CUB pool that contains images on which the pre-trained LLAVA-v1.5-7B and MINIGPT4-vicuna-7B fail the binary habitat classification task. This filtering produces challenging edit requests where the target answer corrects a pre-edit failure, rather than merely confirming knowledge that the model already expresses reliably.

Evaluation data construction. Each retained image is converted into a standard multimodal editing instance with one edit request, two generality probes, and two locality probes. The construction of each field is as follows.

Reliability. The edit request uses the retained failure image and the source prompt

```
What type of bird is in the image? Choose
from: land, water.
```

with the corrected habitat label as the new target answer. This field tests whether an editor can correct the pre-edit failure on the original edit input.

Text-generality. Text-generality keeps the same edit image and target answer, but replaces the source prompt with a paraphrased binary habitat question. We randomly sample from the following paraphrase templates:

```
1. Based on its physical features, would
you classify this bird as a water bird
or a land bird?

2. Does this bird primarily inhabit
aquatic environments or terrestrial
ones? Answer with 'water' or 'land'.

3. Identify the habitat type of this
bird: water or land?

4. Is the bird shown here adapted for
life in water or on land?
```

These paraphrases ask for the same water-vs.-land habitat decision without mentioning the background label. This field tests whether the edit transfers across textual reformulations of the same visual fact.

Image-generality. Image-generality keeps the source prompt and target answer, but replaces the edit image with a different image of the same bird species from the full CUB pool. Because the species is unchanged, the desired habitat answer remains unchanged. The replacement image may preserve or change the background type, which lets us test whether the edited answer follows the bird species rather than the surrounding scene. This field is the main diagnostic probe: a robust edit should follow the bird species across image changes rather than switch according to the new background.

Text-locality. Following MMEdit (Cheng et al., 2023), text-locality uses unrelated question-answer pairs from NQ (Kwiatkowski et al., 2019). They test whether the edit changes unrelated text-side knowledge.

Image-locality. Following MMEdit (Cheng et al., 2023), image-locality uses unrelated multimodal question-answer examples from OK-VQA (Marino et al., 2019). They test whether the edit changes unrelated multimodal behavior.

Splits and group distribution. The final WATERBIRD-EDIT annotation file contains 1,136 edit instances over 90 bird species, split into 636 training instances and 500 test instances. The main diagnostic results use the 500 test instances as a sequential editing stream. The test edit images remain strongly target-context imbalanced, which reflects the spurious correlations present in the filtered failure pool. By contrast, the same-species image-rephrase set gives substantially broader coverage of the four target-context groups. Table 5 reports the group counts used in the test split.

Image Set	WW	WL	LW	LL
Edit images	18	190	107	185
Image-rephrase images	115	93	133	159

Table 5: WATERBIRD-EDIT test-split group counts. Each group is written as (target habitat, background), where W and L denote water and land. The image-rephrase set is used for the group Image-Generality and PG diagnostics in Section 2.

CUB species-level habitat labels. Table 6 lists the binary habitat label assigned to each of the 200 CUB species classes. All images of the same species have the same y label.

B Additional Implementation Details for VICE

Image-summary generation prompt. Image summaries are generated offline and cached before editing. Following the main text, we instantiate the frozen summary model G as a pretrained MLLM before editing; in our preprocessing script, G is implemented with LLAVA-v1.5-7B. Given an image and its prompt, the summary generator uses the following prompt. The method section writes this compactly as $G(i, x)$; in the implementation, the `{text}` field is instantiated with the prompt only, without the target answer.

```
USER: <image>
Given the image and the text prompt,
write a concise visual summary for
knowledge editing. The summary must
identify the target-related visual
evidence, describe other salient
objects or attributes that could act as
distractors, and briefly contrast the
target evidence against those
distractors. Do not answer the question
directly. Do not simply copy the text
prompt.
Return exactly three lines in the
following format:
Target Evidence: ...
Distractors: ...
Contrast: ...
Text Prompt: {text}
ASSISTANT:
```

Generation is deterministic with greedy decoding and a maximum of 128 new tokens. The returned summary is stored together with the corresponding image path and prompt, and is reused when constructing the evidence-summary demonstrations.

Scope-head training data. The scope-discriminative head h_ψ is trained offline for each dataset using only the training split. For each training edit instance, we form one scope group. The reliability example is used as the anchor, and its text-rephrase and image-rephrase examples are treated as in-scope queries with the same group label. Text-locality and image-locality examples are treated as out-of-scope queries; each locality example is assigned its own unique label so it is not used as a positive example for any edit anchor.

Scope features and projection head. We encode each query (i, x) with CLIP-ViT-B/32 (`openai/clip-vit-base-patch32`). The text prompt is encoded with the CLIP text encoder. If an image is available, it is encoded with the CLIP image encoder; for text-only locality examples, we use a zero image feature. Both text and image features are ℓ_2-normalized, concatenated, and kept fixed during scope-head training. The learned head is a two-layer MLP,

$$h_\psi(\cdot) = \mathrm{Linear}_{d\to 256}(\mathrm{ReLU}(\mathrm{Linear}_{d\to d}(\cdot))), \tag{10}$$

where d is the concatenated CLIP feature dimension. The projected features are ℓ_2-normalized before computing retrieval similarities.

Optimization objective. Training minibatches are sampled at the edit-group level: each batch contains 32 edit groups by default and includes all scope examples from those groups. Within a minibatch, reliability examples provide the candidate anchors. For non-anchor in-scope examples, we use the anchor retrieval loss in Section 3, which is a cross-entropy objective over the minibatch anchors. For locality examples, we use the margin loss in Section 3, which penalizes cases where the maximum similarity to any minibatch anchor exceeds m. The full objective is the same as in Section 3:

$$\mathcal{L}_{\text{scope}} = \mathcal{L}_{\text{ret}} + \mathcal{L}_{\text{margin}}. \tag{11}$$

Unless otherwise specified, we use AdamW for 80 epochs with learning rate 10^{-3}, temperature $T = 0.07$, projection dimension 256, and margin $m = 0.5$. After training, the dataset-specific head is frozen and used only for memory retrieval and threshold routing.

C Complete Experimental Results

This section provides the full lifelong editing results that complement the condensed main table in Section 4. In addition to the 1, 100, and 1000 edit settings reported in the main text, we include the intermediate 10-edit setting for both LLaVA-v1.5 and MiniGPT4.

Table 6: CUB species-level habitat labels used by WATERBIRD-EDIT.

Class	Class	Class	Class
001 Black footed Albatross (water)	**002** Laysan Albatross (water)	**003** Sooty Albatross (water)	**004** Groove billed Ani (land)
005 Crested Auklet (water)	**006** Least Auklet (water)	**007** Parakeet Auklet (water)	**008** Rhinoceros Auklet (water)
009 Brewer Blackbird (land)	**010** Red winged Blackbird (land)	**011** Rusty Blackbird (land)	**012** Yellow headed Blackbird (land)
013 Bobolink (land)	**014** Indigo Bunting (land)	**015** Lazuli Bunting (land)	**016** Painted Bunting (land)
017 Cardinal (land)	**018** Spotted Catbird (land)	**019** Gray Catbird (land)	**020** Yellow breasted Chat (land)
021 Eastern Towhee (water)	**022** Chuck will Widow (land)	**023** Brandt Cormorant (water)	**024** Red faced Cormorant (water)
025 Pelagic Cormorant (water)	**026** Bronzed Cowbird (land)	**027** Shiny Cowbird (land)	**028** Brown Creeper (land)
029 American Crow (land)	**030** Fish Crow (land)	**031** Black billed Cuckoo (land)	**032** Mangrove Cuckoo (land)
033 Yellow billed Cuckoo (land)	**034** Gray crowned Rosy Finch (land)	**035** Purple Finch (land)	**036** Northern Flicker (land)
037 Acadian Flycatcher (land)	**038** Great Crested Flycatcher (land)	**039** Least Flycatcher (land)	**040** Olive sided Flycatcher (land)
041 Scissor tailed Flycatcher (land)	**042** Vermilion Flycatcher (land)	**043** Yellow bellied Flycatcher (land)	**044** Frigatebird (water)
045 Northern Fulmar (water)	**046** Gadwall (water)	**047** American Goldfinch (land)	**048** European Goldfinch (land)
049 Boat tailed Grackle (land)	**050** Eared Grebe (water)	**051** Horned Grebe (water)	**052** Pied billed Grebe (water)
053 Western Grebe (water)	**054** Blue Grosbeak (land)	**055** Evening Grosbeak (land)	**056** Pine Grosbeak (land)
057 Rose breasted Grosbeak (land)	**058** Pigeon Guillemot (water)	**059** California Gull (water)	**060** Glaucous winged Gull (water)
061 Heermann Gull (water)	**062** Herring Gull (water)	**063** Ivory Gull (water)	**064** Ring billed Gull (water)
065 Slaty backed Gull (water)	**066** Western Gull (water)	**067** Anna Hummingbird (land)	**068** Ruby throated Hummingbird (land)
069 Rufous Hummingbird (land)	**070** Green Violetear (land)	**071** Long tailed Jaeger (water)	**072** Pomarine Jaeger (water)
073 Blue Jay (land)	**074** Florida Jay (land)	**075** Green Jay (land)	**076** Dark eyed Junco (land)
077 Tropical Kingbird (land)	**078** Gray Kingbird (land)	**079** Belted Kingfisher (land)	**080** Green Kingfisher (land)
081 Pied Kingfisher (land)	**082** Ringed Kingfisher (land)	**083** White breasted Kingfisher (land)	**084** Red legged Kittiwake (water)
085 Horned Lark (land)	**086** Pacific Loon (water)	**087** Mallard (water)	**088** Western Meadowlark (water)
089 Hooded Merganser (water)	**090** Red breasted Merganser (water)	**091** Mockingbird (land)	**092** Nighthawk (land)
093 Clark Nutcracker (land)	**094** White breasted Nuthatch (land)	**095** Baltimore Oriole (land)	**096** Hooded Oriole (land)
097 Orchard Oriole (land)	**098** Scott Oriole (land)	**099** Ovenbird (land)	**100** Brown Pelican (water)
101 White Pelican (water)	**102** Western Wood Pewee (water)	**103** Sayornis (land)	**104** American Pipit (land)
105 Whip poor Will (land)	**106** Horned Puffin (water)	**107** Common Raven (land)	**108** White necked Raven (land)
109 American Redstart (land)	**110** Geococcyx (land)	**111** Loggerhead Shrike (land)	**112** Great Grey Shrike (land)
113 Baird Sparrow (land)	**114** Black throated Sparrow (land)	**115** Brewer Sparrow (land)	**116** Chipping Sparrow (land)
117 Clay colored Sparrow (land)	**118** House Sparrow (land)	**119** Field Sparrow (land)	**120** Fox Sparrow (land)
121 Grasshopper Sparrow (land)	**122** Harris Sparrow (land)	**123** Henslow Sparrow (land)	**124** Le Conte Sparrow (land)
125 Lincoln Sparrow (land)	**126** Nelson Sharp tailed Sparrow (land)	**127** Savannah Sparrow (land)	**128** Seaside Sparrow (land)
129 Song Sparrow (land)	**130** Tree Sparrow (land)	**131** Vesper Sparrow (land)	**132** White crowned Sparrow (land)
133 White throated Sparrow (land)	**134** Cape Glossy Starling (land)	**135** Bank Swallow (land)	**136** Barn Swallow (land)
137 Cliff Swallow (land)	**138** Tree Swallow (land)	**139** Scarlet Tanager (land)	**140** Summer Tanager (land)
141 Artic Tern (water)	**142** Black Tern (water)	**143** Caspian Tern (water)	**144** Common Tern (water)
145 Elegant Tern (water)	**146** Forsters Tern (water)	**147** Least Tern (water)	**148** Green tailed Towhee (land)
149 Brown Thrasher (land)	**150** Sage Thrasher (land)	**151** Black capped Vireo (land)	**152** Blue headed Vireo (land)
153 Philadelphia Vireo (land)	**154** Red eyed Vireo (land)	**155** Warbling Vireo (land)	**156** White eyed Vireo (land)
157 Yellow throated Vireo (land)	**158** Bay breasted Warbler (land)	**159** Black and white Warbler (land)	**160** Black throated Blue Warbler (land)
161 Blue winged Warbler (land)	**162** Canada Warbler (land)	**163** Cape May Warbler (land)	**164** Cerulean Warbler (land)
165 Chestnut sided Warbler (land)	**166** Golden winged Warbler (land)	**167** Hooded Warbler (land)	**168** Kentucky Warbler (land)
169 Magnolia Warbler (land)	**170** Mourning Warbler (land)	**171** Myrtle Warbler (land)	**172** Nashville Warbler (land)
173 Orange crowned Warbler (land)	**174** Palm Warbler (land)	**175** Pine Warbler (land)	**176** Prairie Warbler (land)
177 Prothonotary Warbler (land)	**178** Swainson Warbler (land)	**179** Tennessee Warbler (land)	**180** Wilson Warbler (land)
181 Worm eating Warbler (land)	**182** Yellow Warbler (land)	**183** Northern Waterthrush (land)	**184** Louisiana Waterthrush (land)
185 Bohemian Waxwing (land)	**186** Cedar Waxwing (land)	**187** American Three toed Woodpecker (land)	**188** Pileated Woodpecker (land)
189 Red bellied Woodpecker (land)	**190** Red cockaded Woodpecker (land)	**191** Red headed Woodpecker (land)	**192** Downy Woodpecker (land)
193 Bewick Wren (land)	**194** Cactus Wren (land)	**195** Carolina Wren (land)	**196** House Wren (land)
197 Marsh Wren (land)	**198** Rock Wren (land)	**199** Winter Wren (land)	**200** Common Yellowthroat (land)

		EVQA					EIC					VLKEB				
# Edits	**Method**	Rel	T-Gen	I-Gen	T-Loc	I-Loc	Rel	T-Gen	I-Gen	T-Loc	I-Loc	Rel	T-Gen	I-Gen	T-Loc	I-Loc
1	FT-VL	100.00	91.03	95.38	95.29	90.53	**99.93**	**94.65**	**97.34**	91.63	82.16	**100.00**	**98.74**	**99.92**	76.13	84.15
	TP	33.27	31.33	29.06	96.77	98.83	57.49	57.48	53.91	95.41	96.52	50.34	52.43	50.27	92.59	96.58
	MEND	91.23	90.05	91.29	91.02	90.22	92.82	91.81	90.59	96.38	93.69	92.13	91.28	90.22	89.19	90.13
	SERAC	89.33	83.72	84.97	82.05	23.78	88.18	81.03	85.61	84.01	28.58	89.77	89.11	87.92	66.68	14.20
	LiveEdit	94.28	**94.51**	88.01	100.00	100.00	82.16	81.01	78.27	100.00	100.00	96.43	95.22	93.72	100.00	**100.00**
	VICE	**100.00**	87.33	**97.57**	**100.00**	**100.00**	98.91	92.76	79.61	**100.00**	**100.00**	99.98	98.27	96.72	**100.00**	99.98
10	FT-VL	91.05	81.96	85.11	83.89	68.17	85.74	81.72	**79.73**	77.89	77.60	86.31	82.23	84.95	66.64	80.62
	TP	32.98	30.90	28.77	91.08	96.14	55.10	55.04	51.65	75.19	88.18	49.59	52.15	49.70	87.02	93.70
	MEND	3.58	3.55	3.53	2.10	1.26	66.49	64.77	58.42	87.25	86.15	0.18	0.24	0.05	0.03	0.19
	SERAC	88.09	83.40	83.57	64.91	15.50	56.57	56.88	52.62	59.96	14.66	81.55	74.49	80.24	54.71	13.15
	LiveEdit	93.79	**93.21**	86.42	100.00	100.00	81.93	80.80	75.55	100.00	100.00	95.54	94.52	91.25	100.00	**100.00**
	VICE	**100.00**	87.18	**97.53**	**100.00**	**100.00**	**98.91**	**92.76**	77.92	**100.00**	**100.00**	**99.98**	**98.27**	**97.14**	**100.00**	99.96
100	FT-VL	76.44	67.30	67.76	58.04	46.18	73.11	69.78	65.46	52.21	58.69	75.91	72.95	74.89	56.36	70.29
	TP	32.60	30.74	28.65	83.13	89.68	50.31	50.67	46.83	34.35	63.78	48.26	50.78	48.20	64.18	79.15
	MEND	2.22	2.20	2.21	0.21	0.62	56.83	56.89	53.07	87.63	84.64	0.56	0.58	0.66	0.18	0.07
	SERAC	88.08	81.53	82.48	62.13	12.90	53.35	53.70	49.41	48.04	17.25	72.25	62.43	70.68	53.73	13.69
	LiveEdit	93.54	**92.34**	85.89	100.00	99.31	80.81	78.77	63.52	100.00	**100.00**	94.56	90.65	89.56	100.00	**100.00**
	VICE	**100.00**	87.15	**97.50**	**100.00**	**99.82**	**98.91**	**92.76**	**73.97**	**100.00**	99.95	**99.98**	**98.27**	**97.14**	**100.00**	99.85
1000	FT-VL	60.21	50.76	51.03	24.37	29.48	64.21	59.43	57.09	33.39	38.30	65.30	62.67	63.87	33.89	37.32
	TP	32.82	30.48	28.55	42.12	60.28	33.75	36.27	32.34	10.47	26.22	34.23	34.74	34.44	27.11	45.52
	MEND	0.04	0.05	0.05	0.08	0.09	54.39	54.14	50.99	83.87	80.60	0.03	0.05	0.07	0.06	0.08
	SERAC	85.57	75.58	82.01	62.46	15.69	52.93	53.44	49.01	49.91	16.65	60.93	56.49	60.06	52.94	15.04
	LiveEdit	92.93	**90.16**	84.30	100.00	96.43	72.80	69.95	57.05	100.00	99.79	92.22	83.97	82.75	100.00	**100.00**
	VICE	**100.00**	87.03	**97.01**	**100.00**	**98.38**	**98.91**	**92.76**	**60.91**	**100.00**	**99.90**	**99.98**	**98.26**	**96.75**	**100.00**	98.32

Table 7: Complete lifelong editing results for LLaVA-v1.5 on EVQA, EIC, and VLKEB. FT-VL, TP, and VICE are from our local runs; MEND, SERAC, and LiveEdit entries are copied from Chen et al. (2025).

		EVQA					EIC					VLKEB				
# Edits	**Method**	Rel	T-Gen	I-Gen	T-Loc	I-Loc	Rel	T-Gen	I-Gen	T-Loc	I-Loc	Rel	T-Gen	I-Gen	T-Loc	I-Loc
1	FT-VL	**100.00**	**99.74**	**99.86**	63.66	35.65	**100.00**	**95.28**	**96.41**	58.21	33.84	**100.00**	**99.75**	**99.95**	56.12	42.03
	TP	30.07	28.55	30.40	89.48	71.50	30.79	32.00	34.41	44.72	28.74	37.17	38.31	37.49	74.56	46.15
	MEND	93.47	91.32	91.22	81.51	74.97	90.28	83.91	79.05	95.92	93.97	90.28	89.18	89.55	93.63	86.69
	SERAC	87.66	69.24	85.96	82.11	24.11	85.09	78.96	81.70	83.36	24.50	89.28	89.06	87.35	64.88	12.72
	LiveEdit	94.90	92.63	92.06	100.00	100.00	86.84	83.82	76.95	100.00	100.00	96.41	95.56	92.85	100.00	**100.00**
	VICE	99.98	88.74	97.35	**100.00**	**100.00**	98.93	93.67	77.85	**100.00**	**100.00**	95.27	82.95	92.05	**100.00**	99.99
10	FT-VL	72.26	64.96	67.67	56.13	30.94	73.47	70.26	67.19	29.40	14.59	75.21	73.90	73.35	48.01	35.70
	TP	18.82	18.45	18.58	38.24	21.89	14.54	14.61	14.49	2.78	1.25	14.66	14.60	13.90	9.60	3.89
	MEND	37.34	31.66	35.52	44.31	37.24	69.98	69.35	59.25	93.39	91.52	57.01	55.39	55.34	72.03	67.66
	SERAC	86.61	67.86	83.14	67.69	14.47	50.72	49.60	48.36	62.87	16.01	79.90	71.80	78.44	52.64	13.08
	LiveEdit	94.06	**91.84**	90.60	100.00	99.92	85.79	83.52	**76.09**	100.00	100.00	**95.75**	**95.13**	91.99	100.00	**100.00**
	VICE	**99.98**	88.54	**97.31**	**100.00**	**100.00**	**98.93**	**93.67**	76.08	**100.00**	**100.00**	95.27	82.95	**92.44**	**100.00**	99.97
100	FT-VL	56.02	49.77	50.35	16.97	10.42	59.24	56.32	52.58	4.13	4.65	64.49	63.10	62.70	21.95	14.10
	TP	6.64	6.16	6.04	1.80	0.61	7.74	7.44	7.23	1.08	0.63	4.93	4.85	4.40	1.57	1.34
	MEND	23.32	20.75	21.42	47.27	42.60	51.92	53.96	48.24	86.02	83.88	38.85	40.02	37.99	72.82	68.89
	SERAC	87.39	66.53	81.70	65.13	14.04	47.41	46.17	45.22	49.37	13.52	69.38	58.31	67.57	51.71	13.72
	LiveEdit	93.23	**91.27**	88.34	100.00	99.61	84.38	83.53	**71.85**	100.00	**100.00**	94.06	**94.62**	89.42	100.00	**100.00**
	VICE	**99.98**	88.49	**97.28**	**100.00**	**99.82**	**98.93**	**93.67**	71.73	**100.00**	99.95	**95.27**	82.95	**92.44**	**100.00**	99.87
1000	FT-VL	36.32	31.16	34.89	5.27	5.86	62.25	57.95	54.73	7.16	11.06	42.64	41.62	41.18	8.55	9.32
	TP	7.70	6.99	7.50	1.70	0.55	8.16	8.26	7.63	0.59	1.46	3.74	3.47	3.57	1.17	2.75
	MEND	31.84	26.60	33.98	43.08	44.90	49.30	51.46	46.70	81.03	82.53	42.35	45.07	42.68	66.98	62.32
	SERAC	84.50	60.36	81.83	63.26	13.40	47.03	45.79	44.82	42.12	15.31	57.02	51.66	56.80	48.48	13.24
	LiveEdit	92.75	**89.81**	85.12	100.00	96.82	81.43	80.34	**62.25**	100.00	99.73	93.41	**89.87**	85.21	100.00	**100.00**
	VICE	**99.98**	88.37	**96.82**	**100.00**	**98.26**	**98.93**	**93.67**	57.31	**100.00**	**99.92**	**95.27**	82.95	**92.14**	**100.00**	98.29

Table 8: Complete lifelong editing results for MiniGPT4 on EVQA, EIC, and VLKEB. FT-VL, TP, and VICE are from our local runs; MEND, SERAC, and LiveEdit entries are copied from Chen et al. (2025).