Evidence Guided Dual Expert Memory with Joint Routing for VLLMs Online Correction

Lintong Zhang **Yearang Lee** **Seong-Whan Lee***
Dept. of Artificial Intelligence, Korea University, Seoul, Korea
{ zhanglintong, yr_lee, sw.lee}@korea.ac.kr

Abstract

Vision–large language models (VLLMs) editing aims to efficiently update model responses while retaining generalization and unrelated knowledge. Recent online editors store previous edits as reusable experts, forming a mixture-of-experts editing memory for scalable VLLM correction. By retrieving relevant experts at inference time, this paradigm avoids repeatedly overwriting model parameters and enables continual reuse of accumulated editing experience. However, existing sample-level residual experts usually encode each edit instance as a whole, entangling target-relevant visual evidence with irrelevant context. This coarse expert construction weakens expert reusability, leads to inaccurate routing, and increases unintended side effects during continual editing. To address this limitation, we reformulate online VLLM editing as evidence guided expert construction and routing. A fine mask identifies edit-relevant visual evidence, from which support-suppress experts are constructed to strengthen target consistent cues and weaken edit conflicting cues. Each expert is stored with routing signals and reliability scores, and later retrieved through joint sparse routing over visual evidence, textual context, and expert quality. Experiments demonstrate consistent improvements in editing reliability, generalization, and interpretability over competitive online editing baselines.

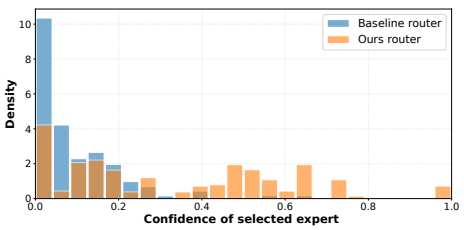
1 Introduction

Vision-language large model (VLLM) editing aims to update specific knowledge or revise erroneous responses without costly retraining. Early studies mainly focus on text-only LLMs, covering factual association localization, transformer memory mass editing, and alignment with edited knowledge [27, 28, 18]. The resulting editing techniques have been applied to knowledge sanitization, privacy protection, security analysis, and bias mitigation [15, 17, 24, 36]. Existing multimodal editing methods further study editable knowledge storage and modification in VLLMs [6, 2, 5], while most existing multimodal editors address one-shot correction rather than continual reuse. To improve scalability in continual editing, retrieval-based methods store previous edit records and retrieve matching edit records during inference instead of repeatedly overwriting model parameters [4, 11]. Among them, LiveEdit [3] further advances lifelong VLLM editing with low-rank experts, where each edit becomes a lightweight corrective expert for future reuse.

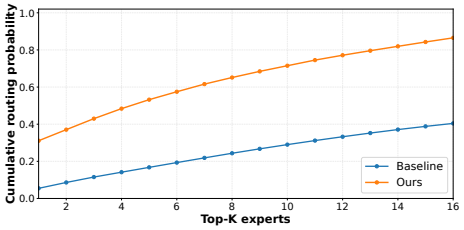
However, existing lifelong VLLM editing methods still suffer from unreliable expert reuse, leading to inconsistent and low-confidence routing decisions. To demonstrate this issue, we analyze the routing patterns of the baseline such as LiveEdit [3]. Fig. 1(a) compares the confidence distribution of the selected experts between the baseline and our method. As shown in Fig. 1(a), the baseline tends to activate more experts with low-confidence scores, indicating that expert selection is uncertain and unreliable compared to our method. The unreliable expert selection in the baseline further leads to dispersed routing distributions, where probability mass is spread across many experts instead of

*Corresponding author

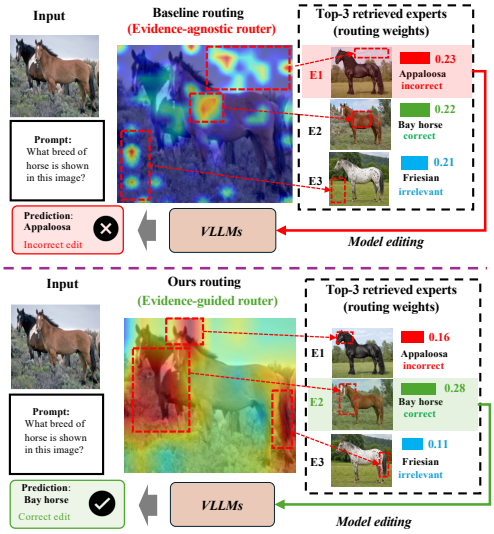
40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Distribution of top-k expert confidence.



(b) Top-k cumulative routing mass.



(c) Diffuse routing mass.

Figure 1: Limitation of expert routing for the baseline. (a) The baseline is distracted by noisy regions and retrieves mismatched experts, causing an incorrect edit toward "*Appaloosa*" instead of "*Bay horse*". Baseline (b) activates low-confidence experts and (c) is diffusely spread across many experts.

concentrating on a few reliable ones. Fig. 1(b) shows the cumulative routing probability with respect to the number of selected experts. As shown in Fig. 1(b), the cumulative routing probability of the baseline grows slowly as more experts are included, representing that routing probability is diffusely distributed across many candidates rather than concentrated on a few reliable experts. This dispersed probability mass indicates that the router lacks confidence in selecting a set of high-quality experts.

To further understand why the baseline selects unreliable experts, we visualize the attention heatmaps and corresponding editing results of an editing example in Fig. 1(c). As shown in Fig. 1(c) top, the baseline routing method is concentrated on irrelevant or weakly related image regions, such as background information. As a result, the router attends to misleading visual evidence and assigns high routing scores to mismatched experts associated with "*Appaloosa*", while the correct expert "*Bay horse*" receives insufficient weight. These unreliable and diffuse routing decisions lead to incorrect editing results.

To address this limitation, we propose **EGDER**, an **E**vidence-**G**uided **D**ual **E**xpert Memory with Joint **R**outing for online VLLM correction. EGDER focuses on decision relevant visual evidence and leverages it to construct evidence guided expert memories for reliable and discriminative routing decisions. (1) We first introduce an Evidence Guided Routing Module that refines standard cross-attention into token level necessity scores for visual tokens. The module then derives a soft mask from the necessity scores to filter evidence regions. Expert generation is performed only within regions critical to the editing target. (2) In Dual Expert Establishment stage, the selected visual evidence is applied to construct paired support-suppress experts. The support expert branch strengthens target consistent cues, and the suppress expert branch weakens hallucination prone cues. (3) We introduce Joint Sparse Routing for inferencetime expert retrieval and fusion. This strategy jointly considers visual evidence quality, prompt consistency, and expert reliability when selecting and weighting experts, thereby improving expert reuse precision and reducing low-confidence expert activation. As shown in Fig. 1(c) bottom, EGDER focuses on correction-relevant visual evidence and performs evidence-aware expert routing, producing accurate editing results.

The main contributions of this paper are as follows:

- We identify a limitation of online VLLM editors: sample-level experts store residual corrections from the whole visual context with edit-irrelevant visual regions and causing noisy high frequency reuse.
- We propose EGDER, an evidence guided dual expert memory framework with joint sparse routing and dual experts to retrieve correct experts.

- Experiments on VLLM editing benchmarks show that EGDER achieves state-of-the-art performance. We further provide a new perspective for evaluation of expert quality.

2 Related Work

2.1 Vision Large Language Models

Vision-Language Large Models (VLLMs) adapt large language models to visual inputs by coupling pretrained vision encoders with language backbones. Representative systems such as Qwen [1], LLaVA [25], and BLIP-2 [23] map visual representations into the language space through lightweight projectors or query-based transformers. This architecture supports image-conditioned understanding and generation [20, 19], yet it also couples visual and textual knowledge within shared representations. As a result, hallucinations and factual errors become difficult to repair after deployment, which motivates model editing methods for targeted VLLM knowledge correction.

2.2 Model Editing

Model editing corrects specific errors or updates knowledge without full retraining. Existing LLM editors fall into three lines: parameter-modification methods such as ROME [27] and MEMIT [28], meta-learning editors [8, 29], and memory-based retrieval editors [18, 11]. Lifelong editing further addresses sequential interference through sparse or expert modules. ELA [26] edits attributed FFN neurons, SEMoE [7] stores knowledge experts for serial edits, ELDER [22] uses mixture-of-LoRA, and MoEEdit [10] studies routing-stable editing for native MoE LLMs. These methods are text-centric or tied to intrinsic MoE architectures, with no explicit treatment of visual evidence selection or cross-modal correction. For VLLMs, MMEdit [6] and VLKEB [13] provide multimodal editing benchmarks, and attribution-based work studies how visual representations influence textual outputs [5, 12, 37, 31]. Recent lifelong VLLM editors, including LiveEdit [3] and CARML [39], use low-rank expert routing or conflict-aware retrieval. They still rely on global sample-level retrieval or output-level signals, leaving fine-grained visual evidence underused during expert construction and reuse. Our method builds evidence-aware support-suppress experts from fine-grained visual cues and reuses them through quality-aware joint sparse routing.

2.3 Mixture-of-Experts

Mixture-of-Experts (MoE) [16, 32] partitions computation across specialized modules, providing a scalable route to larger model capacity with conditional activation. Prior work improves routing and load balancing in large-scale MoE models [21, 9]. Recent studies examine decoder-only MoE training with modified pre-training objectives, including UL2-style objectives [35]. Mixtral [33] adopts token-choice routing and activates two of eight experts, which increases routing flexibility and model performance. Our work repurposes MoE for online VLLM editing rather than pre-training scale-up. Each correction is stored as a lightweight low-rank expert that represents a reusable local editing operator. Mask-guided evidence selection restricts expert construction to edit-relevant visual tokens, and quality-aware joint routing controls when stored experts are reused. This formulation turns MoE into an evidence-conditioned correction memory for continual VLLM editing.

3 Preliminaries

We study online VLLM editing. Let $f_\Theta : \mathcal{X} \times \mathcal{P} \to \mathcal{A}$ denote a VLLM with parameters Θ that maps an image x and a prompt p to $\hat{a} = f_\Theta(x, p)$. Given an edit stream $\mathcal{D}_e = \{(x_n, p_n, a_n)\}_{n=1}^{N_e}$, each sample provides an image, a prompt, and a desired corrected answer. The editing objective covers edited-sample correction, transfer to related inputs, and prediction preservation on unrelated inputs. At editing layer l_e, the n-th sample (x_n, p_n, a_n) yields $v_r^{(n)}$, $p_r^{(n)}$, and $a_r^{(n)}$. In retrieval-based MoE editing, each edit forms a lightweight correction expert in $\mathcal{E}$ rather than updating Θ. The routing memory $\mathcal{M}$ stores retrieval features $R_v^{(n)}$, $R_p^{(n)}$, and $R_q^{(n)}$ for expert selection. During inference, the router matches input representations with $\mathcal{M}$, retrieves candidate experts from $\mathcal{E}$, and assigns sparse weights W. The selected experts produce Δ_h, which updates a_h into a_c for final decoding while Θ remains fixed.

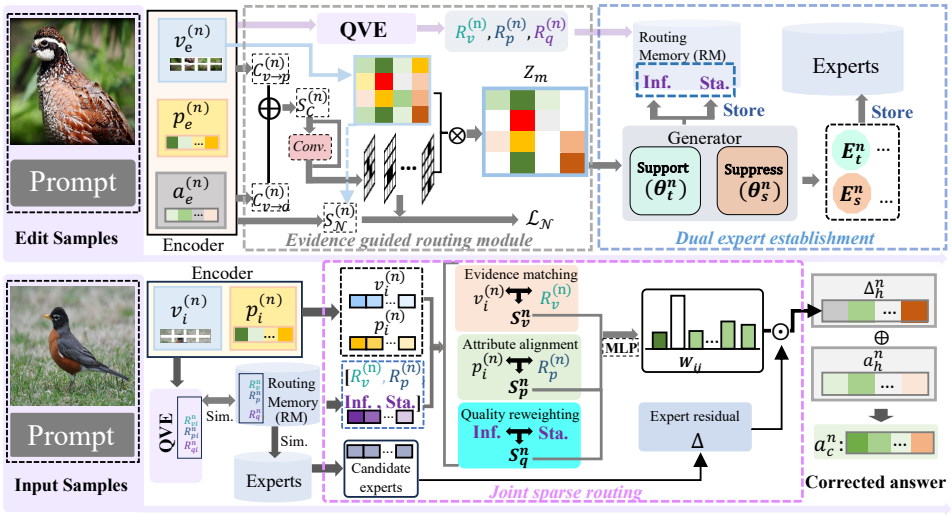


Figure 2: **Overview of EGDER. Top:** During editing, the input image, prompt, and target answer pass through evidence guided routing to filter edit relevant visual regions. The filtered evidence guides paired support and suppress expert construction, with routing records stored for reuse. **Bottom:** During inference, QVE matches input routing features with stored records and retrieves the closest candidate experts. Joint sparse routing scores evidence matching, prompt alignment, and expert quality. The router assigns sparse weights for expert fusion, and the fused residual update revises the hidden answer representation.

4 Method

4.1 Overview

We describe EGDER, an evidence-guided online editing framework for VLLMs. As shown in Fig. 2, EGDER contains three components: evidence guided routing, dual expert establishment, and joint sparse routing. Evidence guided routing filters correction-relevant visual evidence and records routing cues. Dual expert establishment converts the filtered evidence into paired support and suppress experts. Joint sparse routing retrieves and fuses stored experts by considering visual evidence, prompt alignment, and expert quality.

4.2 Evidence Guided Routing Module

First, to construct evidence experts that rectify erroneous model decisions, we identify visual regions that are important to the editing target and build experts from these regions. Given $v_e^{(n)}$, $p_e^{(n)}$, and $a_e^{(n)}$ from the n-th edit sample, where M denotes the number of visual tokens in $v_e^{(n)}$, we first estimate how each visual token correlates with the prompt and the target answer. The two cross-modal evidence scores are computed as:

$$c_{v\to p}^{(n)} = \mathrm{Attn}_{vp}(v_e^{(n)}, p_e^{(n)}) \in \mathbb{R}^M, \qquad c_{v\to a}^{(n)} = \mathrm{Attn}_{va}(v_e^{(n)}, a_e^{(n)}) \in \mathbb{R}^M, \tag{1}$$

where $\mathrm{Attn}_{vp}(\cdot)$ and $\mathrm{Attn}_{va}(\cdot)$ denote cross-modal attention functions from visual tokens to prompt and answer tokens, respectively.

The prompt-conditioned and answer-conditioned evidence scores provide complementary cues for locating edit-relevant visual tokens. The coarse token-level evidence score $S_c^{(n)}$ is computed as:

$$S_c^{(n)} = \rho\left([c_{v\to p}^{(n)}; c_{v\to a}^{(n)}]\right) \in \mathbb{R}^M, \tag{2}$$

where $\rho(\cdot)$ denotes a token-wise linear projection that maps the fused evidence scores to one coarse evidence logit for each visual token.

Since the coarse score contains noisy token responses caused by sparse cross-modal correlations, we further refine it with local smoothing to obtain a fine mask $\mathcal{F}_e^{(n)}$. Specifically, $\mathrm{Conv}_{1d}(\cdot)$

performs local smoothing over neighboring visual tokens, which reduces isolated noisy activations and preserves spatially coherent evidence responses. Meanwhile, to identify tokens that are both visually informative and consistent with the editing target, $S_{\mathcal{N}}^{(n)}$ is computed from the visual representation and the prompt and answer-conditioned evidence scores. These two outputs are formulated as:

$$\mathcal{F}_e^{(n)} = \sigma\left(\mathrm{Conv}_{1d}(S_c^{(n)}) + S_c^{(n)}\right) \in \mathbb{R}^M, \qquad S_{\mathcal{N}}^{(n)} = \sigma\left([v_e^{(n)}; c_{v\to p}^{(n)}; c_{v\to a}^{(n)}]\right) \in \mathbb{R}^M. \quad (3)$$

To make the necessity branch learn stable token-level evidence patterns without directly affecting the fine mask refinement branch, we supervise it using the detached fine mask. Specifically, the necessity branch is trained with $\mathcal{L}_{\mathcal{N}} = \phi(S_{\mathcal{N}}^{(n)}, \xi(\mathcal{F}_e^{(n)}))$, where $\phi(\cdot)$ denotes binary cross-entropy and $\xi(\cdot)$ denotes the stop-gradient operator, which treats $\mathcal{F}_e^{(n)}$ as a fixed supervision target and prevents gradients from updating the fine mask branch.

Once the fine mask is obtained, the masked visual representation and the masked edit representation $Z_m^{(n)}$ is defined as:

$$Z_m^{(n)} = \mathcal{F}_e^{(n)} \odot v_e^{(n)} \in \mathbb{R}^{(M+M_p+M_a)\times d}, \quad (4)$$

where $\odot$ denotes token-wise multiplication with broadcasting over the hidden dimension, and M_p and M_a denote the numbers of prompt tokens and target answer tokens, respectively.

The resulting $Z_m^{(n)}$ is then passed to the dual expert generator. In this way, expert construction is guided by visual evidence while preserving prompt and target answer semantics, reducing the chance of encoding irrelevant visual background into the expert.

4.3 Dual Expert Establishment

Second, to avoid storing coarse residuals that entangle target-supporting evidence with hallucination-related noise, we build a dual expert mechanism from the masked evidence representation $Z_m^{(n)}$. Since $Z_m^{(n)}$ is obtained through necessity-based token filtering and semantic fusion, it provides an evidence-focused basis for expert construction.

An expert generator $\mathcal{G}$ converts $Z_m^{(n)}$ into support and suppress expert parameters:

$$(\theta_t^{(n)}, \theta_s^{(n)}, \mu^{(n)}, \lambda^{(n)}) = \mathcal{G}(Z_m^{(n)}), \quad (5)$$

where $\theta_t^{(n)}$ and $\theta_s^{(n)}$ denote the support and suppress expert parameters, respectively. The scalar $\mu^{(n)}$ controls the overall correction magnitude and is predicted from the evidence representation, while the dual expert generator applies a fixed residual scale to keep the correction conservative. The coefficient $\lambda^{(n)} \in [0, 1]$ balances the two branches and is learned by a lightweight gate network. The support expert strengthens target-consistent correction directions, whereas the suppress expert weakens residual directions associated with the hallucinated prediction.

For later expert reuse, we use the query vision extractor (QVE) to encode the visual, prompt, and answer representations into routing keys:

$$(R_v^{(n)}, R_p^{(n)}, R_q^{(n)}) = \mathrm{QVE}(v_e^{(n)}, p_e^{(n)}, a_e^{(n)}). \quad (6)$$

Finally, the generated parameters are converted into stored operators $E_t^{(n)}$ and $E_s^{(n)}$, and then written into the expert repository with routing metadata:

$$\mathcal{M}^{(n)} = \mathcal{M}^{(n-1)} \cup \{R_v^{(n)}, R_p^{(n)}, R_q^{(n)}, S_n^{(n)}, \mu^{(n)}, \lambda^{(n)}\}, \qquad \mathcal{E}^{(n)} = \mathcal{E}^{(n-1)} \cup \{E_t^{(n)}, E_s^{(n)}\}. \quad (7)$$

The coupled storage makes each edit reusable as both a corrective operator and a memory item with evidence, routing, and quality cues.

4.4 Joint Sparse Routing

Third, to reduce biased expert activation caused by noisy attention signals, we introduce a joint sparse routing module for reliable expert reuse. For each inference sample i, QVE produces $(R_v^{(i)}, R_p^{(i)}, R_q^{(i)})$, which are compared with each stored memory item j. The router jointly evaluates

visual evidence consistency, prompt-level compatibility, and expert reliability, where each stored expert is associated with quality metadata $\mathrm{Inf.}^{(j)}$ and $\mathrm{Sta.}^{(j)}$ for measuring correction influence and reuse stability. Specifically, the router computes three compatibility scores:

$$S_v^{ij} = \psi_v(R_v^{(i)}, R_v^{(j)}), \quad S_p^{ij} = \psi_p(R_p^{(i)}, R_p^{(j)}), \quad S_q^{ij} = \psi_q\left(R_q^{(i)}, R_q^{(j)}, \mathrm{Inf.}^{(j)}, \mathrm{Sta.}^{(j)}\right), \tag{8}$$

where S_v^{ij}, S_p^{ij}, and S_q^{ij} denote evidence matching, prompt alignment, and quality-aware compatibility scores, respectively. The quality-aware score incorporates $\mathrm{Inf.}^{(j)}$ and $\mathrm{Sta.}^{(j)}$ to suppress experts with weak correction effect or unstable reuse behavior.

Candidate selection and weight assignment are then performed using the same joint routing signals:

$$\mathcal{S}_i = \mathrm{TopK}_j\left(S_v^{ij} + S_p^{ij} + S_q^{ij}, K\right), \tag{9}$$

where $\mathcal{S}_i$ denotes the selected expert set for the i-th input. The TopK operation retains the most relevant experts and filters out unreliable candidates before fusion, making expert activation sparse.

To adaptively determine the contribution of each selected expert, we compute the fusion weight by jointly considering routing similarity and expert reliability:

$$W_{ij} = \mathrm{Softmax}_{j\in\mathcal{S}_i}\left(\mathrm{MLP}\left([S_v^{ij}; S_p^{ij}; S_q^{ij}; \mathrm{Inf.}^{(j)}; \mathrm{Sta.}^{(j)}]\right)\right), \tag{10}$$

where W_{ij} denotes the normalized contribution of the j-th selected expert. Experts with stronger evidence consistency, higher influence, and better stability receive larger fusion weights. In this way, joint sparse routing reduces the activation of visually similar but unreliable experts during both expert selection and fusion. Finally, the selected experts are fused into a corrective residual:

$$\Delta_h^{(i)} = \sum_{j\in\mathcal{S}_i} W_{ij}\mu^{(j)}\left[\lambda^{(j)}E_t^{(j)}(a_h^{(i)}) - (1-\lambda^{(j)})E_s^{(j)}(a_h^{(i)})\right], \qquad a_c^{(i)} = a_h^{(i)} + \Delta_h^{(i)}. \tag{11}$$

The corrected representation $a_c^{(i)}$ is passed through the remaining layers to generate the final response.

5 Experiments

5.1 Implementation Details

We evaluate our evidence-guided dual MoE editor on E-VQA [6] and VLKEB [13] datasets using BLIP2-OPT (2.7B) and LLaVA-V1.5 (7B) models. E-VQA provides controlled multimodal edits, while VLKEB contains real-image samples with more complex visual evidence. Following prior protocols [6, 3], we compare with three groups of baselines: general model editors, including FT-L, FT-M, MEND [29], TP [14], SERAC [30], LTE [18], and RECIPE [4]; lifelong expert-based editors, including LEMoE [34] and LiveEdit [3]; and recent MoE-based lifelong LLM editors, such as MoEEdit [10] and ELDER [22], which are text-only but related to expert-based sequential editing. Following the lifelong setting [3], edits are injected sequentially with #Edit=1000. We report Reliability (Rel.), Text Generality (T-Gen.), modal generality (M-Gen.), text locality (T-Loc.), and Average, and attribution map validation metric Ins. [38].

5.2 Performance Comparisons

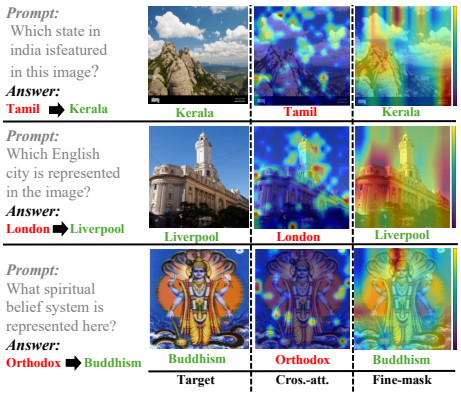
Table 1 shows that our method achieves the best average performance on two datasets. Our method achieves the best performance against recent SOTA methods. The improvement remains pronounced over MoE based online editing methods, including LiveEdit and LEMoE, with gains mainly on Rel., T-Gen., M-Gen., and Ins. Performance gains arise from evidence guided expert construction, support-suppress dual experts, and quality-aware joint sparse routing. The three components jointly construct cleaner experts, select more reliable corrections, and improve lifelong VLLM editing while preserving locality.

5.3 Interpretability Analysis

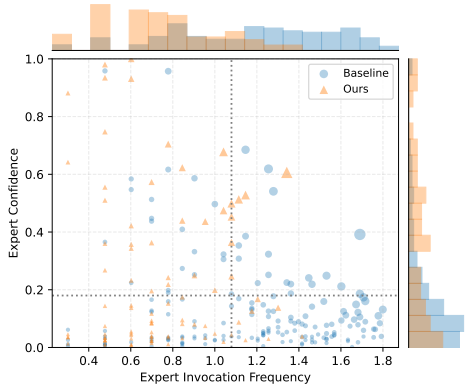
To verify the interpretability of our method from the perspectives of localized editing scope and expert selection, Fig. 3 provides a qualitative analysis of both editing scope and expert routing.

Table 1: Comparison of SOTA methods on E-VQA and VLKEB under two VLLM backbones. All values are reported in percentages (%). Average is computed over Rel., T-Gen., M-Gen., and T-Loc., excluding Ins.

Backbone	Method	E-VQA						VLKEB					
		Rel.	T-Gen.	M-Gen.	T-Loc.	Average	Ins.	Rel.	T-Gen.	M-Gen.	T-Loc.	Average	Ins.
LLaVA-V1.5 (7B)	RECIPE	87.00	76.81	83.09	86.95	83.46	38.56	62.00	56.84	61.50	85.37	66.43	40.63
	FT-L	71.39	59.83	57.41	55.55	61.05	32.47	68.14	66.38	66.98	65.61	66.78	38.21
	LTE	83.93	82.55	81.34	83.97	82.95	37.31	64.51	56.26	64.80	80.85	66.61	39.74
	MEND	0.04	0.05	0.05	0.08	0.06	34.62	0.03	0.05	0.07	0.06	0.05	37.08
	TP	16.56	16.80	15.65	7.28	14.07	36.74	5.46	4.81	5.51	2.77	4.64	37.46
	SERAC	85.57	75.58	82.01	62.46	76.41	35.09	60.93	56.49	60.06	52.94	57.61	38.92
	FT-M	69.57	56.34	44.07	100.00	67.50	33.18	53.41	48.80	43.16	100.00	61.34	37.85
	LEMoE	30.80	25.75	24.32	71.45	38.08	39.12	67.97	61.07	58.16	48.48	58.92	39.31
	LiveEdit	92.93	90.16	84.30	100.00	91.85	39.87	92.22	83.97	82.75	100.00	89.74	41.36
	MoEEdit	93.08	90.02	84.48	99.86	91.86	39.73	92.06	84.12	82.91	99.84	89.73	41.21
	ELDER	88.42	85.67	79.86	96.74	87.67	38.92	86.74	78.92	77.45	97.15	85.07	40.63
	Ours	**95.43**	**93.92**	**88.93**	**100.00**	**94.57**	**41.08**	**95.90**	**93.73**	**89.38**	**100.00**	**94.75**	**41.92**
BLIP2-OPT (2.7B)	RECIPE	84.99	74.20	82.04	96.82	84.51	34.81	54.64	46.54	54.10	94.60	62.47	34.12
	FT-L	45.10	34.62	35.42	48.42	40.89	28.43	55.39	54.34	53.87	50.80	53.60	32.05
	LTE	89.32	82.82	81.51	94.86	87.13	33.47	61.67	51.05	61.60	94.78	67.28	33.24
	MEND	15.84	14.35	17.73	91.74	34.92	30.72	37.22	38.03	37.19	91.49	50.98	31.36
	TP	20.63	15.09	18.41	8.65	15.70	32.95	24.36	24.21	24.25	16.37	22.30	30.58
	SERAC	83.35	70.80	80.32	67.66	75.53	31.38	53.58	45.78	52.42	56.81	52.15	30.94
	FT-M	40.40	31.46	27.85	100.00	49.93	29.16	47.03	49.68	46.99	100.00	60.93	31.72
	LEMoE	19.73	17.34	18.22	72.01	31.83	35.69	34.74	33.43	32.05	55.55	38.94	30.27
	LiveEdit	94.42	91.98	84.65	100.00	92.76	36.54	97.00	91.92	87.53	100.00	94.11	34.61
	MoEEdit	94.58	91.83	84.82	99.82	92.76	36.39	96.84	92.08	87.69	99.83	94.11	34.78
	ELDER	89.72	86.95	79.88	96.86	88.35	35.84	91.56	86.74	82.93	97.02	89.56	33.88
	Ours	**96.36**	**94.74**	**87.28**	**100.00**	**94.60**	**38.46**	**98.02**	**93.68**	**90.16**	**100.00**	**95.47**	**34.93**



(a) Editing scope heatmap



(b) Expert confidence-frequency distribution

Figure 3: Visualization of localized editing and expert routing. The editing scope heatmap shows that the correction focuses on visual evidence, while the expert routing distribution demonstrates that the corresponding edits are selectively routed to a compact set of relevant experts.

Evidence Guided Editing Scope Visualization. Fig. 3(a) visualizes the editing scope heatmaps produced by different methods. The coarse editing scope based on the original cross-attention is relatively dispersed and tends to cover background regions or contextual areas that are correlated with the answer but not necessarily decision-critical. In contrast, our fine mask based localization produces a more compact and clearer editing scope, which is better aligned with the target evidence region. This indicates that our method does not perform coarse intervention over the entire image or broad visual context, but instead constrains the edit to the local visual evidence that truly affects answer correction. Therefore, this result verifies the interpretability of our method in terms of *where to edit*, since it explicitly identifies which visual regions should be used for correction. To quantitatively validate the editing scope, Table 3 compares the fine mask against human-annotated regions, grouped by **Evidence Type (ET)**. We report four localization metrics: **Pointing Game (PG)**, the percentage of samples whose highest-scoring location falls within annotated evidence; **Intersection over Union (IoU)**, the spatial overlap between selected and annotated regions; **Evidence Recall (ER)**, the fraction of annotated evidence covered by the selected regions; and **Irrelevant Mass (IM)**, the proportion of mask weight assigned outside annotated evidence. IoU and ER are evaluated using the top 20% of predicted mask locations. Our method achieves an overall PG of 79.8%, IoU of 0.43, and ER of 0.73.

Table 2: Comparison with MoE-based model editing methods under different VLLM backbones on E-VQA and VLKEB. LE, LA, CD, and Rel. are reported in percentage (%), SE is the number of activated experts, and CT is a unitless concentration score.

Backbone	Method	E-VQA						VLKEB					
		LE↑	LA↓	SE↓	CT↑	CD↑	Rel.↑	LE↑	LA↓	SE↓	CT↑	CD↑	Rel.↑
LLaVA-V1.5 (7B)	LEMoE	10.41	9.58	272	1.52	13.48	30.80	10.91	9.06	246	1.67	14.63	67.97
	ELDER	13.27	8.62	226	1.82	15.74	70.28	12.92	8.48	219	1.88	16.02	75.24
	LiveEdit	15.89	10.08	202	2.03	13.17	82.51	15.42	10.36	196	1.94	12.68	88.43
	MoEEdit	24.29	4.93	226	2.85	18.36	80.27	23.47	3.88	184	2.82	23.58	83.64
	Ours	**28.43**	**1.38**	**74**	**4.96**	**29.24**	**87.43**	**27.86**	**1.56**	**78**	**4.82**	**28.71**	**90.90**
BLIP2-OPT (2.7B)	LEMoE	10.18	9.88	298	1.42	12.31	58.67	10.31	9.61	286	1.48	12.74	70.54
	ELDER	11.32	8.84	236	1.73	14.96	72.15	11.74	8.66	228	1.79	15.31	76.80
	LiveEdit	11.64	9.72	178	1.54	9.86	89.42	12.18	9.40	173	1.66	10.22	84.21
	MoEEdit	18.53	9.12	233	1.65	14.27	87.25	15.26	4.22	215	2.5	19.58	85.31
	Ours	**21.87**	**2.06**	**96**	**3.68**	**22.41**	**92.36**	**22.96**	**1.84**	**90**	**3.90**	**23.34**	**89.57**

Table 3: Localization against human annotations.

ET	PG↑	IoU↑	ER↑	IM↓
Standard	86.3	0.54	**0.78**	0.18
Small Evidence	77.6	0.42	0.67	0.24
Ambiguous Evidence	72.4	0.32	0.61	0.29
Multi-region Evidence	75.1	0.40	0.64	0.27
All	**79.8**	**0.43**	0.73	**0.22**

Table 4: Failure taxonomy and suppression risks.

FT	NC	FS (%)	OR (%)	RR (%)
Wrong mask	30	25.0	3.0	70.0
Wrong retrieval	34	28.3	3.4	76.5
Harmful suppress	15	12.5	1.5	60.0
Expert conflict	20	16.7	2.0	55.0
Overcorrection	12	10.0	1.2	83.3
Other	9	7.5	0.9	11.1
Total	120	100.0	12.0	–

Localization is less accurate for ambiguous evidence, suggesting that uncertain visual cues remain challenging to isolate.

Expert Quality Performance Evaluation. To evaluate the quality of the generated experts during online editing, we construct a set of 200 editing samples and compare our method with existing baselines. We assess five routing-oriented indicators: **Localized Editing (LE)**, **Low-noise Activation (LA)**, **Sparse Experts (SE)**, **Concentration (CT)**, and **Confidence (CD)**. LE measures whether the routing mass is assigned to high-contribution and low-noise experts. Lower LA and SE indicate fewer noisy activations and fewer activated experts per query, while higher CT and CD indicate more concentrated routing distributions and stronger average routing weights. As shown in Table 2, after editing 200 samples, our method consistently achieves higher editing reliability while activating fewer experts and assigning stronger routing weights. The results show that our evidence guided expert construction and joint sparse routing generate cleaner and more reusable experts than competing methods. To further examine the relation between expert confidence and invocation frequency, Fig. 3(b) shows that our method produces more experts in the high-confidence region and substantially fewer frequently invoked experts in the low-confidence region. These results demonstrate that our method learns a compact set of high-quality experts for frequent reuse, thereby reducing the performance degradation caused by repeated activation of low-quality experts.

5.4 Ablation Study

Editing Components. To evaluate the effect of the fine-mask and dual-expert mechanisms on the quality of generated experts, we report the results in Table 5(a). C1 and C2 correspond to two variants with different component configurations. The results show that both the fine-mask module and the dual-expert design consistently improve expert quality, leading to higher LE and CD and lower LA. This indicates that evidence-guided expert construction can reduce noisy activations and disentangle target-supportive correction directions from hallucination-suppressive ones. In addition, we further examine the necessity of the mask and the effect of the loss term $\mathcal{L}_n$. As shown in Table 5(b), the fine mask provides cleaner correction-relevant evidence and achieves the best performance in terms of Rel., M-Gen., and Ins.

Table 5: Ablation studies of our main components on BLIP2-OPT (2.7B).

(a) Expert Construction Ablation

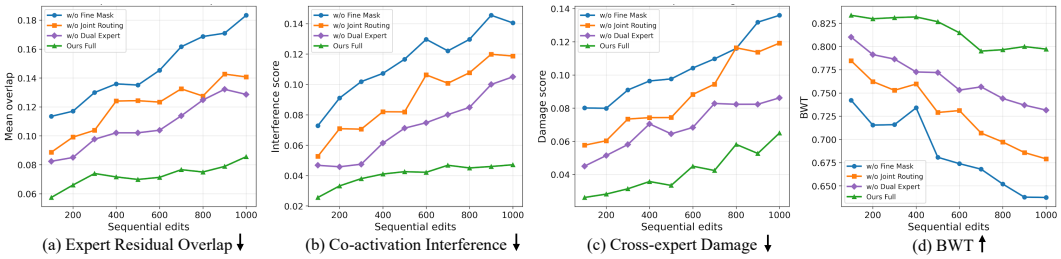
Variant	Expert Construction		Expert Quality		
	Fine-mask	Dual	LE↑	LA↓	CD↑
Base	✗	✗	12.18	9.40	10.22
C1	✓	✗	18.67	4.36	18.14
C2	✗	✓	16.42	6.27	17.06
Full	✓	✓	**22.96**	**1.84**	**23.34**

(b) Mask Quality Ablation

Mask Type	Metrics		
	Rel.↑	M-Gen.↑	Ins.↑
No Mask	94.42	84.65	36.54
Cross-Att. Mask	95.02	85.72	37.48
Fine Mask w/o $\mathcal{L}_n$	94.37	85.06	36.91
Fine Mask	**96.36**	**87.28**	**38.46**

(c) Joint Routing Ablation

Variant	Routing Components			Metrics		
	EM	AL	QR	Rel.↑	M-Gen.↑	Ins.↑
Base	✗	✗	✗	94.42	84.65	36.54
R1	✗	✓	✓	95.22	86.71	37.36
R2	✓	✗	✓	95.61	86.98	37.82
R3	✓	✓	✗	96.12	87.14	38.17
Full	✓	✓	✓	**96.36**	**87.28**	**38.46**



(a) Expert Residual Overlap ↓ (b) Co-activation Interference ↓ (c) Cross-expert Damage ↓ (d) BWT ↑

Figure 4: Expert interference analysis under sequential editing. Our Full reduces residual overlap, co-activation interference, and cross-expert damage, while maintaining higher BWT than variants without the fine mask, joint routing, or dual-expert design.

Routing Strategy. To verify the effectiveness of joint sparse routing, we conduct ablation studies on its three components: evidence matching (EM), alignment modeling (AL), and quality-aware routing (QR). R1, R2, and R3 denote three variants with different routing-component configurations. As shown in Table 5(c), incorporating EM, AL, and QR progressively improves Rel., M-Gen., and Ins. This demonstrates that reliable expert reuse requires jointly modeling visual evidence, prompt-level semantic consistency, and expert reliability.

Expert Interference Analysis. To further evaluate how the fine mask, dual experts, and joint routing affect lifelong editing stability, Fig. 4 reports expert interference as the number of sequential edits increases. The variant without the fine mask shows the highest expert residual overlap and cross-expert damage, indicating that broader visual representations introduce more entangled residual directions. Removing joint routing leads to higher co-activation interference, suggesting that unreliable routing tends to activate incompatible experts together. The variant without dual experts also exhibits stronger interference than the full model, showing that separating supportive and suppressive correction directions helps reduce expert conflict. In contrast, Our Full consistently achieves the lowest residual overlap, co-activation interference, and cross-expert damage, while maintaining the highest BWT. These results demonstrate that fine-mask evidence selection, dual-expert construction, and joint sparse routing jointly suppress expert interference and improve long-term editing stability.

To examine the remaining limitations of our method, we analyze the major sources of editing failures and their recoverability. Table 4 categorizes failures by **Failure Type (FT)**. **Number of Cases (NC)** counts failures in each category, **Failure Share (FS)** measures their proportion among all failures, **Overall Rate (OR)** indicates their frequency among evaluated edits, and **Recovery Rate (RR)** measures the proportion of recovered failures within each category. The results identify incorrect evidence localization and expert retrieval as the primary failure sources, while harmful suppression and expert conflicts remain challenges despite the reduced interference observed in Fig. 4.

Computational Complexity Analysis. To assess the computational efficiency of our framework, we examine the overhead introduced by individual components and compare resource consumption with existing editing methods. Table 6 reports the editing and inference costs under different component configurations, showing that our method achieves improved editing performance with modest computational overhead. Furthermore, Table 7 compares parameter requirements, editing memory, routing latency, and inference efficiency against LTE [18] and RECIPE [4]. The results demonstrate the advantages of our method in parameter efficiency and expert routing, while highlighting the trade-offs between memory consumption and end-to-end latency.

Table 6: Computational cost and performance analysis under different component settings on LLaVA-V1.5 (7B) and BLIP2-OPT. The first row reports absolute values of the base editor, while subsequent rows report relative changes (Δ) with respect to the base editor. WT and PM denote wall-clock time and peak memory, respectively. Each cell is reported as LLaVA-V1.5 (7B) / BLIP2-OPT.

Method	**Edit-time Cost**		**Inference-time Cost**		**Performance**	
	WT (ms)	**PM** (GB)	**WT** (ms)	**PM** (GB)	**Rel.**↑	**Ins.**↑
Base Editor	188.72 / 146.50	40.64 / 34.52	194.86 / 150.27	21.34 / 14.68	94.42 / 97.00	38.46 / 34.93
+ Fine-mask	+5.20 / +4.10	+0.04 / +0.03	+1.30 / +0.90	+0.02 / +0.01	+0.72 / +0.65	+1.71 / +1.44
+ Dual Experts	+4.75 / +3.80	+0.05 / +0.04	+1.05 / +0.78	+0.02 / +0.02	+0.98 / +0.87	+0.55 / +0.92
+ Fine-mask + Dual	+8.66 / +6.75	+0.07 / +0.05	+2.10 / +1.55	+0.03 / +0.02	+2.05 / +1.22	+1.76 / +2.00
+ EM	+0.80 / +0.62	+0.01 / +0.01	+3.60 / +2.80	+0.03 / +0.02	+1.10 / +0.76	+0.35 / +0.26
+ AL	+0.65 / +0.50	+0.01 / +0.01	+2.40 / +1.90	+0.02 / +0.02	+1.47 / +1.05	+0.42 / +0.55
+ QR	+1.00 / +0.78	+0.02 / +0.01	+2.95 / +2.35	+0.03 / +0.02	+1.34 / +1.00	+0.51 / +1.03
Ours-Full	+10.95 / +8.50	+0.10 / +0.07	+6.46 / +5.40	+0.06 / +0.05	+1.94 / +1.50	+2.01 / +2.54

Table 7: Budget comparison relative to EGDER.

Method	**Type**	**Params (M)**	**Edit Mem. (MB)**	**Mem./Edit (KB)**	**Routing (ms)**	**Edit WT (ms)**	**Infer. WT (ms)**	**Infer. PM (GB)**
LTE	Text-only	2651.68	53.762	550.53	9.70	217.09	135.84	14.655
RECIPE	Retrieval-based	231.64	4.537	46.46	7.30	7.43	36.08	15.048
Ours	MoE-based	126.62	7.812	80.00	2.40	135.80	103.50	14.708

6 Conclusion

In this paper, we propose an evidence-aware online editing method for VLLMs. Our method constructs experts from decision-critical visual evidence through a mask-guided evidence selector, models supportive and suppressive correction directions with dual experts, and retrieves reliable experts using quality-aware joint sparse routing. Extensive experiments show that our method improves editing reliability and generalization while providing more interpretable correction behavior. In addition, we provide a detailed analysis of the quality of experts generated during the editing process from multiple perspectives.

Limitations and Future Work. While our method effectively leverages evidence-guided masking for reliable expert construction and routing, it may be less effective in scenarios where errors stem from complex reasoning or long-range cross-modal dependencies beyond localized visual evidence. In addition, as the expert repository grows, long-term editing may incur increased storage and retrieval overhead, leading to the need for more scalable expert management strategies.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University); No. RS-2024-00457882, AI Research Hub Project; No. RS-2025-02304828, Artificial Intelligence Star Fellowship Support Program to Nurture the Best Talents; and No. RS-2026-25507543, Development of AI Co-Scientist based on Scientific Causal World Model).

Competing Interests. The authors declare no competing interests.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[2] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 7400–7426, 2024.

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

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

[5] Qizhou Chen, Taolin Zhang, Chengyu Wang, Xiaofeng He, Dakan Wang, and Tingting Liu. Attribution analysis meets model editing: Advancing knowledge correction in vision language models with visedit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2168–2176, 2025.

[6] Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 13877–13888, 2023.

[7] YuJu Cheng, Yu-Chu Yu, Kai-Po Chang, and Yu-Chiang Frank Wang. Serial lifelong editing via mixture of knowledge experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30888–30903, 2025.

[8] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, 2021.

[9] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[10] Yupu Gu, Rongzhe Wei, Andy Zhu, and Pan Li. Moeedit: Efficient and routing-stable knowledge editing for mixture-of-experts llms. *arXiv preprint arXiv:2602.10965*, 2026.

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

[12] Jung-Ho Hong, Ho-Joong Kim, Kyu-Sung Jeon, and Seong-Whan Lee. Comprehensive information bottleneck for unveiling universal attribution to interpret vision transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25166–25175. IEEE, 2025.

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

[14] Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. Transformer-patcher: One mistake worth one neuron. *arXiv preprint arXiv:2301.09785*, 2023.

[15] Yoichi Ishibashi and Hidetoshi Shimodaira. Knowledge sanitization of large language models. *arXiv preprint arXiv:2309.11852*, 2023.

[16] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[17] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, 2023.

[18] Yuxin Jiang, Yufei Wang, Chuhan Wu, Wanjun Zhong, Xingshan Zeng, Jiahui Gao, Liangyou Li, Xin Jiang, Lifeng Shang, Ruiming Tang, et al. Learning to edit: Aligning llms with knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4689–4705, 2024.

[19] Yearang Lee, Ho-Joong Kim, and Seong-Whan Lee. Text-infused attention and foreground-aware modeling for zero-shot temporal action detection. *Advances in Neural Information Processing Systems*, 37:9864–9884, 2024.

[20] Yearang Lee, Ho-Joong Kim, and Seong-Whan Lee. TF-CADE: Foreground-concentrated text-video alignment for zero-shot temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2843–2852, 2026.

[21] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[22] Jiaang Li, Quan Wang, Zhongnan Wang, Yongdong Zhang, and Zhendong Mao. Elder: Enhancing lifelong model editing with mixture-of-lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24440–24448, 2025.

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

[24] Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. *arXiv preprint arXiv:2403.13355*, 2024.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916, 2023.

[26] Jinzhe Liu, Junshu Sun, Shufan Shen, Chenxue Yang, and Shuhui Wang. Edit less, achieve more: Dynamic sparse neuron masking for lifelong knowledge editing in llms. *arXiv preprint arXiv:2510.22139*, 2025.

[27] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.

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

[29] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

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

[31] Woo-Jeoung Nam and Seong-Whan Lee. Illuminating salient contributions in neuron activation with attribution equilibrium. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1120–1131, 2024.

[32] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[33] MA Team et al. Mixtral of experts: A high quality sparse mixture-of-experts. *Mistral AI Blog. Accessed: December*, 18:2023, 2023.

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

[35] Fuzhao Xue, Zian Zheng, Yao Fu, Jinjie Ni, Zangwei Zheng, Wangchunshu Zhou, and Yang You. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*, 2024.

[36] Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, 2023.

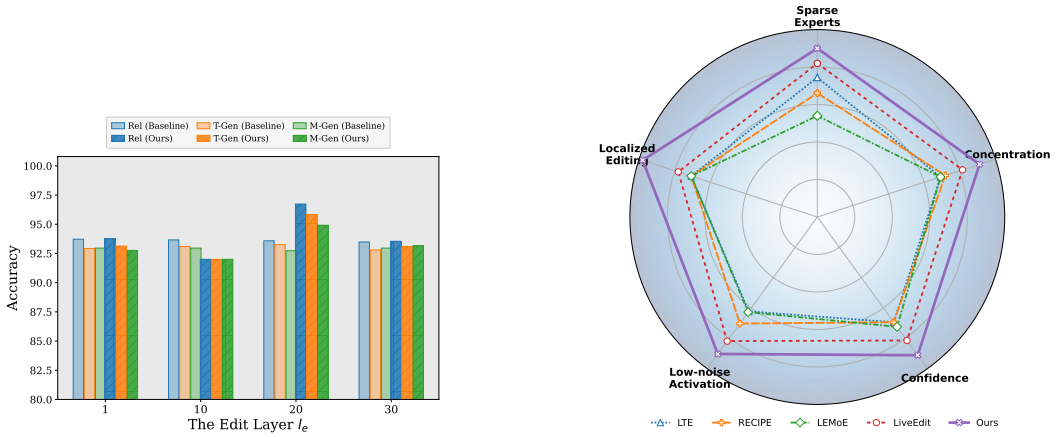
[37] Lintong Zhang, Kang Yin, and Seong-Whan Lee. Semantic prioritization in visual counterfactual explanations with weighted segmentation and auto-adaptive region selection. *Neural Networks*, 184:107097, 2025.

[38] Lintong Zhang, Kang Yin, and Seong-Whan Lee. Towards fine-grained interpretability: Counterfactual explanations for misclassification with saliency partition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30053–30062, 2025.

[39] Qiang Zhang, Fanrui Zhang, Jiawei Liu, Ming Hu, Junjun He, and Zheng-Jun Zha. Reliable lifelong multimodal editing: Conflict-aware retrieval meets multi-level guidance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Evidence Guided Dual Expert Memory with Joint Routing for VLLM Online Correction - Appendix

Editing Layer and Expert Quality Analysis. Fig. A further analyzes the effect of editing layer selection and the quality of generated experts. As shown in Fig. A(a), editing performance varies across attached layers, indicating that the position of expert injection affects correction effectiveness. The results show that shallow layers provide limited improvement, while middle-to-late layers achieve stronger Rel. and T-Gen. performance. In particular, the attached layer around $l_e = 20$ obtains the best overall results, suggesting that this layer provides a better trade-off between visual-textual evidence integration and answer-level correction. Fig. A(b) compares different methods from multiple expert-quality perspectives. Compared with the baselines, our method achieves stronger localized editing, lower noisy activation, higher confidence, better concentration, and sparser expert activation. These results indicate that EGDER not only improves editing performance through proper layer selection, but also constructs cleaner and more reliable experts for selective reuse during online editing.



(a) Editing performance across layers

(b) Expert quality evaluation

Figure A: Analysis of editing layer and expert quality. (a) Editing performance across attached layers, showing the influence of layer selection on editing effectiveness. (b) Expert quality evaluation, showing that our method produces cleaner, more confident, and more concentrated expert routing.

A Routing-oriented Indicators

Given N editing samples, we denote the fine-grained visual evidence mask of the i-th sample as $m_i = \{m_{i,t}\}_{t=1}^{M}$ and the corresponding edit activation intensity on each visual token as $a_i = \{a_{i,t}\}_{t=1}^{M}$. The routing weight assigned to the k-th expert is denoted as $\alpha_{i,k}$, where $\sum_{k=1}^{K} \alpha_{i,k} = 1$.

Localized Editing (LE). LE measures how strongly the editing activation is concentrated on decision-critical visual evidence:

$$\mathrm{LE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{t=1}^{M} m_{i,t} a_{i,t}}{\sum_{t=1}^{M} m_{i,t} + \epsilon}. \tag{12}$$

A higher LE indicates that the generated experts mainly act on evidence-relevant visual tokens.

Low-noise Activation (LA). LA measures the average editing activation outside the selected visual evidence region:

$$\mathrm{LA} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{t=1}^{M} (1 - m_{i,t}) a_{i,t}}{\sum_{t=1}^{M} (1 - m_{i,t}) + \epsilon}. \tag{13}$$

A lower LA indicates fewer noisy activations on irrelevant visual tokens.

Sparse Experts (SE). SE measures the average number of activated experts during routing:

$$\mathrm{SE} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}\left(\alpha_{i,k} > \tau\right), \tag{14}$$

where τ is a small activation threshold. A lower SE indicates sparser expert reuse.

Concentration (CT). CT measures how concentrated the routing distribution is over the expert repository:

$$\mathrm{CT} = \frac{1}{N} \sum_{i=1}^{N} \left(\log K + \sum_{k=1}^{K} \alpha_{i,k} \log(\alpha_{i,k} + \epsilon) \right). \tag{15}$$

This is equivalent to $\log K - H(\alpha_i)$, where $H(\alpha_i)$ is the routing entropy. A higher CT indicates a more concentrated routing distribution.

Confidence (CD). CD measures the margin between the most activated expert and the second most activated expert:

$$\mathrm{CD} = \frac{1}{N} \sum_{i=1}^{N} \left(\alpha_{i,(1)} - \alpha_{i,(2)} \right), \tag{16}$$

where $\alpha_{i,(1)}$ and $\alpha_{i,(2)}$ denote the largest and second-largest routing weights for the i-th sample. A higher CD indicates more confident expert selection.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have described the main contributions of our paper and itemized them.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included the limitations in the conclusion Sec. 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include the theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included the detailed settings in the experiment part Sec. 5.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release the code after the paper is published.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have included the experiment setting in Sec. 5.1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not contain the statistical bar. However, following previous approaches [3], we reported Reliability (Rel.), Text Generality (T-Gen.), modal generality (M-Gen.), text locality (T-Loc.), and Average.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the computational cost analysis in Table. 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have ensured that our research fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper does not contain societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not contain such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the all datasets and previous works.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We will release the code after the paper is published.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Although vision large language models were used in this study, they were not incorporated as an original or unconventional part of the main method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.