BEYOND BINARY EDITS: ROBUST MULTIMODAL KNOWLEDGE EDITING WITH ADVERSARIAL SUBSPACE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) need efficient mechanisms to update knowledge without degrading existing capabilities. While intrinsic multimodal knowledge editing achieves strong reliability and locality, it often exhibits limited generality, failing to propagate edits across semantically equivalent visual and linguistic variations. This issue arises from the lack of explicit semantic supervision, rigid editing scopes, and biased anchoring to individual samples in high-dimensional multimodal spaces. We address robust intrinsic multimodal knowledge editing by explicitly targeting generalization. We formalize robustness through knowledge units that group semantically equivalent multimodal inputs and define generality as consistent predictions within each unit. To expose fragile semantic regions, we introduce Latent Adversarial Robustification (LAR), which generates adversarial yet semantically coherent variants in the joint latent space. We further propose Rank-Constrained Subspace Learning (RCSL), enforcing low-rank alignment of adversarial representations at the edit layer via a singular value–based objective. The resulting framework, Adversarial Subspace AlignMent (ASAM), is evaluated as a plug-and-play module on existing editors. Extensive analysis demonstrates the effectiveness of ASAM empirically.

1 INTRODUCTION

Multimodal large language models (MLLMs) Li et al. (2024a); Bai et al. (2025b) rapidly advance generative AI across applications due to their superior cross-modal comprehension. Knowledge Editing (KE) Wang et al. (2024c) has gained prominence as an efficient means to update knowledge and correct errors in pre-trained models, particularly for MLLMs where editing cross-modal structures enables precise, localized model updates. Existing editing methods for MLLMs fall into two types Pan et al. (2024), *external knowledge resorting* Mitchell et al. (2022); Zheng et al. (2023) and *intrinsic knowledge editing* Mitchell et al. (2021); Huang et al. (2023). The former retrieves edited information from external memories, while the latter directly modifies internal parameters.

While external knowledge resorting retrieves contextual information from auxiliary memories, intrinsic knowledge editing directly updates model parameters to internalize new knowledge, yielding more persistent and robust edits by eliminating dependence on runtime retrieval. Existing intrinsic editing Pan et al. (2024); Guo et al. (2025) achieve *high reliability* on the edited facts and *strong locality* for out-of-scope knowledge, but they often exhibit **limited generality**, i.e., failing to propagate updates to semantically analogous inputs, particularly under varied visual or linguistic rephrasings of the original edit. Therefore, in this paper, we focus on addressing the performance gap in intrinsic multimodal editing by preserving reliability and locality while significantly improving generality, enabling robust and semantically adaptive editing across diverse cross-modal contexts. However, solving the observed gap is non-trivial due to the following challenges, as illustrated in Fig. 1.

CH1. Lack of explicit and scalable supervision for semantic generalization in high-dimensional multimodal spaces. Generalization in multimodal editing cannot be reliably enforced through a simple supervised objective. Although rephrased inputs may appear semantically equivalent at the linguistic level or in high-level cross-attention patterns, their corresponding joint vision–language representations often reside in distant regions of a highly curved manifold. Existing methods Wang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

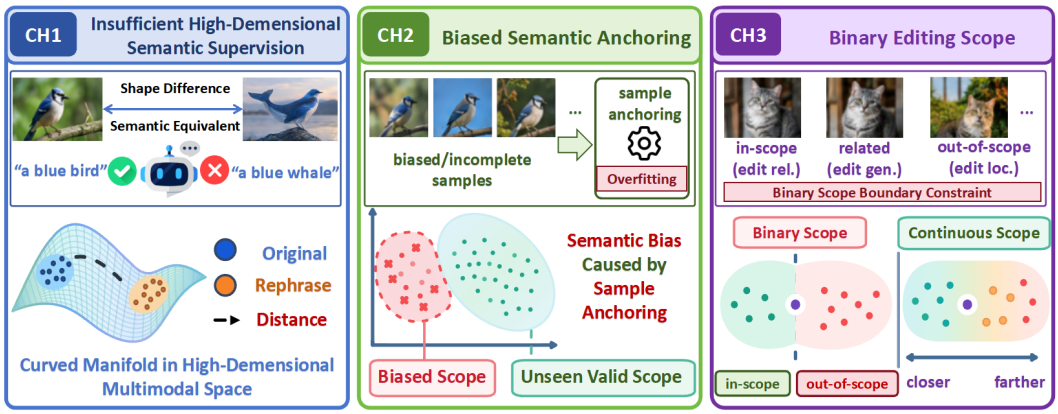


Figure 1: **Three challenges that limit generality in multimodal knowledge editing. (CH1)** A localized parameter update receives insufficient high-dimensional supervision. **(CH2)** Anchoring the update to biased or incomplete samples overfits spurious attributes and leaves valid variants outside the edited support. **(CH3)** Binary in-/out-of-scope partitions ignore graded semantic relatedness, so the editor cannot propagate an edit continuously around its semantic core.

et al. (2024b); Fang et al. (2024) perform a localized deformation of model parameters around limited editing samples without generalization constraints, which induce only a narrow semantic influence region, rendering the edited knowledge brittle to distributional variations induced by changes in visual composition, background context, or attribute phrasing.

CH2. Unreliable semantic anchoring under biased editing samples. For a single editing intent, multimodal samples often exhibit biased or incomplete distributions in the semantic space, making them unreliable references for generalization. Previous methods Guo et al. (2025) anchor parameter updates to individual samples, causing the edited knowledge to absorb incidental visual patterns or linguistic artifacts. Such sample-centric anchoring encourages overfitting to spurious attributes and restricts the generalization boundary. A mechanism is needed to abstract beyond imperfect sample realizations and capture the shared semantic space underlying diverse perturbations. Otherwise, editing is confined to the biased support of observed samples, limiting robustness to rephrased inputs.

CH3. Binary editing scopes obscure graded semantic relatedness under multimodal shifts. Most intrinsic editing methods Mitchell et al. (2021); Pan et al. (2024) partition samples into *in-scope* and *out-of-scope* sets, implicitly imposing a static boundary. This binary formulation oversimplifies the continuous nature of semantic relatedness in multimodal inputs, providing no representation-level mechanism to model graded semantic proximity. Under multimodal distribution shifts, such a rigid scope offers no principled way to determine how far an edit should propagate along meaningful semantic directions. The challenge is therefore to extend edits continuously around their semantic core while maintaining a controlled transition to unchanged model behavior.

To tackle these challenges, we propose an Adversarial Subspace AlignMent based framework (ASAM), which internalizes edited knowledge as a consistent semantic concept across equivalent multimodal variations. ASAM comprises three representation-level components. For **CH1**, Latent Adversarial Robustification (LAR) generates diverse, semantically coherent variants in the joint vision–language latent space without external supervision. For **CH2**, Rank-Constrained Subspace Learning (RCSL) aligns these variants in a low-rank subspace, capturing invariant semantics without anchoring to a single realization. For **CH3**, Asymmetric Gradient Flow continuously expands the edit around its semantic core while detaching the original representation, avoiding rigid editing boundaries. Together, they embed edited knowledge into a low-rank semantic structure that generalizes across multimodal shifts while preserving locality.

Our contributions can be summarized as: (1) We propose an adversarial subspace alignment framework for robust multimodal knowledge editing. (2) We introduce latent adversarial robustification to generate diverse latent variants for semantic supervision. (3) We develop rank-constrained subspace learning with asymmetric gradient flow to organize semantic variants and control edit propagation. (4) Extensive experiments on MMEdit and MMKE-Bench show that ASAM improves generality across heterogeneous MLLMs and editing paradigms while largely preserving reliability and locality.

2 Related Work

Knowledge Editing. Knowledge Editing (KE) modifies factual knowledge in foundation models while preserving unrelated semantics. Methods fall into *external knowledge resorting* Bi et al. (2024); Cohen et al. (2023); Zheng et al. (2023); Gu et al. (2024); Mitchell et al. (2022); Hartvigsen et al. (2023); Wang et al. (2024b), which rely on retrieval and runtime augmentation, and *intrinsic editing* Zhang et al. (2024c;b); Jiang et al. (2025); De Cao et al. (2021); Mitchell et al. (2021); Hase et al. (2023); Yu et al. (2023); Tan et al. (2024); Meng et al. (2022a;b); Dong et al. (2022); Li et al. (2024c); Ma et al. (2024); Hernandez et al. (2024); Tamayo et al. (2024); Fang et al. (2024); Deng et al. (2025); Pan et al. (2025), which modify model parameters via hypernetworks or locate-then-edit strategies for persistent, localized updates. Multimodal extensions (UniKE Pan et al. (2024), BalancEdit Guo et al. (2025), ODEdit Su et al. (2026)) address cross-modal entanglement but generalization remains limited.

Multimodal Language Models. Multimodal LLMs (MLLMs) Zhu et al. (2023); Li et al. (2023); Lin et al. (2024); Dai et al. (2023) unify visual-linguistic representations, enabling perception, reasoning, and generation. Efficient KE is critical to correct outdated or biased knowledge Li et al. (2024b); Zeng et al. (2024), yet entangled representations cause edits to overfit instance-specific inputs, limiting generalization Cheng et al. (2023); Huang et al. (2024); Zhang et al. (2024a); Du et al. (2025).

Detailed discussions of related work are presented in Appendix C.

3 Problem Formulation

Multi-modal Knowledge Editing. Let $f_\theta : \mathcal{X}_v \times \mathcal{X}_t \to \mathcal{Y}$ denote a pre-trained multi-modal large language model parameterized by θ, where $\mathcal{X}_v$ and $\mathcal{X}_t$ represent the visual and textual input spaces, respectively. The goal of Multi-modal Knowledge Editing is to modify f_θ such that new factual knowledge is incorporated, while the model's behavior on unrelated inputs remains invariant. Formally, we have the definition of the knowledge unit as follows:

Definition 3.1 (Knowledge Unit)**.** A *knowledge unit* $\mathcal{K}_i$ is defined as a set of semantically equivalent input pairs $\mathcal{K}_i = \{(x_v^{(j)}, x_t^{(j)})\}_{j=1}^{m_i}$, all of which correspond to a unique ground-truth target $y_i^* \in \mathcal{Y}$.

Robust Knowledge Editing. Conventional knowledge editing aims to obtain updated parameters θ' with an edit request $(x_v, x_t, y^*) \in \mathcal{K}_i$, such that

$$f_{\theta'}(x_v, x_t) = y^*. \tag{1}$$

However, robust knowledge editing requires that the update generalizes beyond a single input sample.

Definition 3.2 (Robust Knowledge Editing)**.** An edit to $\mathcal{K}_i$ is *robust* if the updated model $f_{\theta'}$ satisfies

$$f_{\theta'}(x_v, x_t) = y_i^*, \quad \forall (x_v, x_t) \in \mathcal{K}_i, \tag{2}$$

while preserving the original predictions on $\mathcal{K} \setminus \mathcal{K}_i$.

We define data pairs $(x_v, x_t) \in \mathcal{K}_i$ as *in-scope knowledge* for knowledge unit $\mathcal{K}_i$, and pairs $(x_v, x_t) \notin \mathcal{K}_i$ as *out-of-scope knowledge*. Throughout, $f_\theta(x_v, x_t)$ denotes the model's discrete prediction, obtained as $\arg\max_y f_\theta(y \mid x_v, x_t)$ when f_θ is viewed as a conditional distribution over $\mathcal{Y}$. Let $\mathcal{D}_{\text{edit}}$ be the set of edit requests on which the modification is applied. We evaluate editing with three primary metrics following Zhang et al. (2024d):

$$\begin{cases} \text{Rel.} = \mathbb{E}_{(x_v, x_t, y^*) \sim \mathcal{D}_{\text{edit}}} \left[\mathbb{I}\{ f_{\theta'}(x_v, x_t) = y^* \} \right] \\ \text{Loc.} = \mathbb{E}_{(x_v, x_t) \sim \mathcal{K} \setminus \mathcal{K}_i} \left[\mathbb{I}\{ f_{\theta'}(y \mid x_v, x_t) = f_\theta(y \mid x_v, x_t) \} \right] \\ \text{Gen.} = \mathbb{E}_{(x_v, x_t, y^*) \sim \mathcal{K}_i} \left[\mathbb{I}\{ f_{\theta'}(x_v, x_t) = y^* \} \right] \end{cases} \tag{3}$$

Rel. evaluates the edit success on the specific query, Loc. verifies the preservation of unrelated knowledge, and generality Gen. measures generalization to equivalent instances.

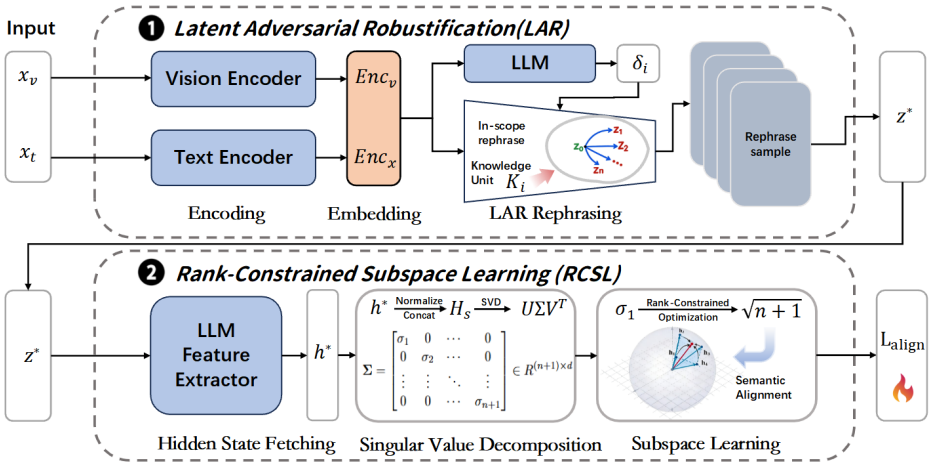


Figure 2: The overall framework of ASAM, which consists of two key modules. ❶ **LAR.** Given multimodal inputs, LAR perturbs input embeddings along LLM-guided gradients to generate semantically consistent rephrases. ❷ **RCSL.** Using these rephrases, RCSL applies SVD-based subspace learning to align editing-layer outputs, enforcing semantic consistency across variants.

4 METHODOLOGY

In this section, we introduce Latent Adversarial Robustification (LAR) to generate semantic variants (Sec. 4.1). Rank-Constrained Subspace Learning (RCSL) is proposed to enforce representation-level consistency across these variants (Sec. 4.2). We then clarify the optimization objective with Asymmetric Gradient Flow proposed to mitigate rigid binary editing scopes (Sec. 4.3).

4.1 LATENT ADVERSARIAL ROBUSTIFICATION (LAR)

To evaluate and enforce the generality risk $\mathcal{R}_{\text{gen}}$, each knowledge unit $\mathcal{K}_i$ requires diverse semantic variants. However, conventional augmentation or manual paraphrasing is costly and limited to human priors, failing to expose model-specific sensitivities or *blind spots*. To address this, we propose *Latent Adversarial Robustification (LAR)*, a gradient-driven, model-internal method that constructs adversarial semantic neighborhoods for robust knowledge editing.

Adversarial Variant Generation. The core idea of LAR is to systematically perturb the input embeddings along directions that maximally challenge the current model parameters, thereby simulating extreme yet semantically coherent variations within the knowledge unit. Let $e_v = \text{Enc}_v(x_v)$, $e_t = \text{Enc}_t(x_t)$ denote the latent embeddings of the visual and textual components of an edit request, respectively. We concatenate them into a joint representation $z = [e_v, e_t]$. To generate challenging variants for $\mathcal{K}_i$, we first compute the cross-entropy loss $\mathcal{L}_{ce}(f_\theta(z), y^*)$ via a forward pass. Unlike standard training, LAR maximizes this loss to induce adversarial perturbations. Following Fu & Barez (2025), we compute the gradient with respect to the latent input:

$$g_z = \nabla_z \mathcal{L}_{ce}(f_\theta(z), y^*). \tag{4}$$

Definition 4.1 (Adversarial Latent Variant)**.** An adversarial variant z_i^* is obtained by solving$z_i^* = z + \delta_i^*, \quad \delta_i^* = \arg\max_{\|\delta\|_p \le \epsilon} \mathcal{L}_{\text{ce}}(f_\theta(z+\delta), y^*)$, where ϵ constrains semantic deviation.

Guided by this gradient, we generate n adversarial variants $z_i^* = [e_v + \delta_{v,i}, e_t + \delta_{t,i}]$ by solving the following inner maximization problem:

$$\delta_i^* = \arg\max_{\|\delta_i\| \le \epsilon} \mathcal{L}_{ce}(f_\theta(z+\delta_i), y^*). \tag{5}$$

In practice, δ_i^* is approximated via single-step projected gradient ascent with randomized initialization, yielding a set $\{z_1^*, \ldots, z_n^*\}$.

Training Objective for Robust Knowledge Editing. By generating variants that oppose standard training updates, LAR exposes the model to boundary cases around the edited knowledge. The

resulting robust editing objective aggregates losses over all n adversarial variants:

$$\mathcal{L}_{\text{robust}} = \mathcal{L}_{\text{align}}\left(\{f_{\theta'}(z_i^*)\}_{i=1}^n\right). \tag{6}$$

This formulation expands a single edit point into a continuous adversarial region in latent space. By enforcing prediction consistency under worst-case perturbations, it encourages a smoother loss landscape, thereby internalizing the edited knowledge as robust, generalized representations rather than brittle instance-level mappings.

4.2 Rank-Constrained Subspace Learning (RCSL)

To overcome the brittleness of traditional knowledge editing in multimodal settings, we propose the *Rank-Constrained Subspace Learning (RCSL)*. Its main insight lies in two parts: 1) Robust generalization requires learning a continuous, adaptive boundary rather than relying on a static and binary in/out-of-scope assumption. And 2) the edit must be anchored to an invariant semantic core that persists across input variations, not to potentially biased individual samples. RCSL operationalizes this by enforcing that all latent variants of an edit converge within a low-rank representation subspace. This ensures the model update captures the essential, shared concept, leading to consistent and reliable behavior on paraphrased inputs.

Formulation. Given an original edit request and its adversarial variants $\{z_0, z_1, \ldots, z_n\}$ generated by LAR (with z_0 denoting the original sample), each input is independently forwarded through the MLLM. Let $h_i \in \mathbb{R}^d$ denote the hidden state extracted from a predefined semantic feature extractor ϕ, which is composed of the edit layer L and the early layers of the language model for input z_i. The edit layer is the sole target of parameter modification during the editing process, formally,

$$h_i := f_\phi(z_i) = f_\phi(z + \delta_i). \tag{7}$$

Our objective is to globally align all samples symmetrically, while preserving the original computational graph for other editing objectives such as reliability and locality.

Rank-Constrained Alignment. To formalize the connection between semantic alignment and low-rank structure, we leverage the geometric properties of the Gram matrix. Let $\mathbf{H}_s = [\hat{\mathbf{h}}_0, \hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_n]^\top \in \mathbb{R}^{(n+1)\times d}$ denote the matrix of ℓ_2-normalized hidden states, where $\hat{\mathbf{h}}_i = \mathbf{h}_i/\|\mathbf{h}_i\|_2$ for $i \in \{0, 1, \ldots, n\}$. The corresponding Gram matrix $\mathbf{G} = \mathbf{H}_s\mathbf{H}_s^\top \in \mathbb{R}^{(n+1)\times(n+1)}$ captures the pairwise inner products between all normalized hidden states.

The entries in $\mathbf{G}$ can represent the cosine similarities between representations as well, thereby characterizing their geometric alignment Cicchetti et al. (2025); Liu et al. (2026; 2025). From Definition 3.2, the robustness objective indicates that all adversarial variants produce identical representations:

$$\hat{\mathbf{h}}_i = \hat{\mathbf{h}}_j, \quad \forall i, j \in \{0, 1, \ldots, n\}. \tag{8}$$

Therefore, achieving this perfect alignment is equivalent to requiring $\mathbf{G} = \mathbf{1}\mathbf{1}^\top$, where $\mathbf{1} \in \mathbb{R}^{n+1}$ denotes the all-ones vector. Notably, matrix $\mathbf{G}$ has rank one, establishing a direct connection between semantic consistency and rank constraints. We formalize this observation in the following proposition:

Proposition 4.2 (Rank-Constrained Alignment)**.** *Let $\{\hat{\mathbf{h}}_0, \hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_n\}$ denote the ℓ_2-normalized hidden states from the editing layer, and define the Gram matrix $\mathbf{G} \in \mathbb{R}^{(n+1)\times(n+1)}$ with entries $\mathbf{G}_{i,j} = \langle \hat{\mathbf{h}}_i, \hat{\mathbf{h}}_j \rangle$. Then the following statements are equivalent: 1) All hidden states are perfectly aligned, i.e., $\hat{\mathbf{h}}_i = \hat{\mathbf{h}}_j \; \forall i, j \in \{0, 1, \ldots, n\}$, 2) $\mathbf{G}_{i,j} = 1 \; \forall i, j \in \{0, 1, \ldots, n\}$, and 3) $\operatorname{rank}(\mathbf{G}) = 1$.*

Remark. This rank-1 structure captures the desired invariance property: despite adversarial perturbations in the input space, the editing layer is expected to encode a shared underlying semantic concept. This naturally leads to the objective of robust knowledge editing: for any samples (x_v, x_t) belonging to the same knowledge unit $\mathcal{K}$, the edited output y^* should remain consistent.

4.3 Asymmetric Gradient Flow for Optimization

To operationalize Proposition 4.2, we leverage the singular value decomposition (SVD) of the hidden state matrix $\mathbf{H}_s$. Specifically, we compute the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{n+1}$ of $\mathbf{H}_s$ as: $\text{SVD}(\mathbf{H}_s) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{n+1})$. Wherein, $\mathbf{U} \in \mathbb{R}^{(n+1)\times(n+1)}, \mathbf{V} \in \mathbb{R}^{d\times d}$ are

orthogonal matrices. The dominance of the leading singular value σ_1 relative to the others serves as a quantitative measure of alignment quality: when $\sigma_1 \gg \sigma_j$ for $j > 1$, the representations are nearly rank-1 and thus well-aligned. We treat the singular values as unnormalized logits and define a temperature-scaled softmax objective for subspace alignment:

$$\mathcal{L}_{\text{align}} = -\log\left(\frac{\exp(\sigma_1/\tau)}{\sum_{j=1}^{n+1}\exp(\sigma_j/\tau)}\right), \tag{9}$$

where $\tau > 0$ is a temperature hyperparameter that controls the sharpness of the alignment constraint. Minimizing $\mathcal{L}_{\text{align}}$ explicitly suppresses secondary singular directions $\{\sigma_j\}_{j>1}$, thereby encouraging all adversarial representations toward a single shared semantic axis within the editing layer.

Our method asymmetrically treats gradient flow. The original hidden state $\hat{\mathbf{h}}_0$ acts as a semantic anchor and is *detached* from $\mathcal{L}_{\text{align}}$, preserving its graph for other objectives. All adversarial variants $\{\hat{\mathbf{h}}_i\}_{i=1}^n$ propagate gradients through $\mathcal{L}_{\text{align}}$, updating only the editing layer parameters θ_L to align diverse inputs with the original semantics:

$$\nabla_{\theta_L}\mathcal{L}_{\text{align}} = \sum_{i=1}^{n}\frac{\partial\mathcal{L}_{\text{align}}}{\partial\hat{\mathbf{h}}_i}\frac{\partial\hat{\mathbf{h}}_i}{\partial\theta_L}, \tag{10}$$

where the summation excludes $i = 0$. This asymmetric gradient flow ensures that robustness is achieved through adaptive parameter updates rather than distortion of the original representation.

The overall optimization objective for robust knowledge editing is formally represented as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{loc}} + \beta\mathcal{L}_{\text{align}}, \tag{11}$$

where $\beta > 0$ balances task-specific fidelity and robustness on the representation level. Reliability and locality losses $\mathcal{L}_{\text{rel}}$ and $\mathcal{L}_{\text{loc}}$ follow standard knowledge editing formulations, ensuring accurate and localized edits. The alignment loss $\mathcal{L}_{\text{align}}$ (Eq. 9) acts on hidden states from adversarial variants $\{z_i\}_{i=1}^n$ generated by LAR, aligning their representations along a shared semantic direction.

Synergies. The core robustness of ASAM stems from the joint optimization of Latent Adversarial Robustification (LAR) and Rank-Constrained Subspace Learning (RCSL). LAR acts as an end-to-end probe, directly uncovering high-dimensional, machine-centric vulnerabilities in the latent space without the computational overhead of external LLMs or diffusion models. RCSL processes these LAR-generated variants, constraining the edited knowledge within a stable low-rank region to resolve the exposed inconsistencies. Together, this tight coupling of adversarial variant generation and rank-constrained alignment ensures that edits generalize seamlessly across challenging multimodal inputs while preserving precise semantic consistency. Our ablation experiments further validate this synergistic effect (Sec. 5.3).

5 Experiments

5.1 Experimental Setups

Benchmarks. Following prior work (Pan et al., 2024), we evaluate on the MMEdit benchmark (Cheng et al., 2023), which comprises two subtasks: Editing VQA (E-VQA) and Editing Image Caption (E-IC). We use four MLLM backbones: BLIP2-OPT (Li et al., 2023) and MiniGPT-4 (Zhu et al., 2023) as primary models, together with Qwen2-VL (Wang et al., 2024a) and LLaVA-OneVision (Li et al., 2024a) to assess architecture-level generality. To further test whether the same representation-level objective transfers to diverse visual knowledge, we additionally evaluate on MMKE-Bench (Du et al., 2025), which organizes edits into three visual-knowledge types: *entity* (public factual associations), *user* (personalized or private associations), and *visual* (appearance-centric attributes). On this benchmark, we use Qwen3-VL-4B (Bai et al., 2025a) and Gemma-4 E4B (Team et al., 2026) as backbones. Architectural details of all models are in Appendix A.1. Unless otherwise specified, performance is assessed via Reliability, Generality, and Locality (T-Loc. and M-Loc.).

Baselines & Implementation Details. We consider 7 baselines across four types, including fine-tuning (FT), MEND (Mitchell et al., 2021), IKE (Zheng et al., 2023), SERAC (Mitchell et al.,

Table 1: **Performance comparison on the MMEdit benchmark.** We report Reliability (Rel.), Generality (Gen.), and textual and multimodal Locality (T-Loc. and M-Loc.) on Editing VQA (E-VQA) and Editing Image Caption (E-IC).

		E-VQA				E-IC			
MLLM	**Method**	**Rel.↑**	**Gen.↑**	**T-Loc.↑**	**M-Loc.↑**	**Rel.↑**	**Gen.↑**	**T-Loc.↑**	**M-Loc.↑**
BLIP2-OPT	WISE	99.00	84.80	100.00	8.99	100.00	82.54	99.84	10.28
	$\text{ASAM}_{\text{WISE}}$	**99.00**	**93.97**	**100.00**	**9.73**	**100.00**	**86.22**	**100.00**	**10.74**
	MEND	95.22	94.17	99.53	92.35	76.10	60.30	98.18	**79.16**
	$\text{ASAM}_{\text{MEND}}$	**95.99**	**94.65**	**99.66**	**93.80**	**77.40**	**62.60**	**98.29**	79.07
	T-Patcher	96.43	87.43	99.93	**22.61**	94.74	70.07	99.93	**37.61**
	$\text{ASAM}_{\text{T-Patcher}}$	**97.38**	**90.42**	**99.93**	20.41	**95.05**	**71.43**	**99.93**	34.83
	UniKE	**100.00**	**100.00**	74.39	7.04	**100.00**	86.51	**100.00**	10.23
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	**100.00**	**81.77**	**12.42**	**100.00**	**94.97**	99.57	**10.77**
MiniGPT-4	WISE	100.00	98.86	100.00	56.48	95.68	83.01	100.00	**89.80**
	$\text{ASAM}_{\text{WISE}}$	**100.00**	**99.00**	**100.00**	**56.75**	**95.84**	**83.36**	**100.00**	88.49
	MEND	96.13	95.41	99.61	**93.92**	82.00	59.60	**99.17**	87.17
	$\text{ASAM}_{\text{MEND}}$	**96.60**	**95.80**	**99.62**	93.60	**82.80**	**61.40**	99.08	**87.64**
	T-Patcher	56.79	55.73	99.79	**75.66**	63.88	58.18	**99.90**	**78.22**
	$\text{ASAM}_{\text{T-Patcher}}$	**68.27**	**66.77**	**99.92**	74.61	**75.83**	**65.86**	99.87	74.43
	UniKE	**100.00**	**99.75**	80.90	14.70	**100.00**	96.35	84.68	19.38
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	99.67	**95.59**	**18.88**	99.66	**99.54**	**96.57**	**19.87**
Qwen2-VL	WISE	100.00	95.32	**84.37**	11.61	99.92	93.61	98.48	**11.15**
	$\text{ASAM}_{\text{WISE}}$	**100.00**	**96.69**	80.02	**11.80**	**99.99**	**94.78**	**99.48**	10.39
	MEND	74.16	76.59	96.51	71.18	88.73	**89.48**	96.12	**66.98**
	$\text{ASAM}_{\text{MEND}}$	**74.16**	**77.78**	**96.51**	**71.18**	**90.50**	87.62	**97.03**	65.46
	T-Patcher	78.79	56.03	**98.51**	**68.62**	73.80	57.39	**98.84**	**70.69**
	$\text{ASAM}_{\text{T-Patcher}}$	**87.83**	**83.17**	**99.03**	46.86	**82.39**	**63.28**	98.74	60.53
	UniKE	**100.00**	**97.51**	41.99	10.17	**100.00**	96.64	**58.89**	**13.21**
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	97.17	**43.51**	**10.58**	**100.00**	**96.83**	57.96	13.09
LLaVA-OneVision	WISE	**100.00**	89.06	**97.83**	15.35	100.00	92.32	97.93	11.63
	$\text{ASAM}_{\text{WISE}}$	99.00	**90.87**	97.11	**15.70**	**100.00**	**93.71**	**99.54**	**12.49**
	MEND	82.67	85.65	97.32	77.22	71.24	75.78	**97.48**	71.39
	$\text{ASAM}_{\text{MEND}}$	**83.17**	**88.90**	**97.74**	**79.44**	**93.37**	**90.68**	96.97	**72.43**
	T-Patcher	79.56	56.62	**99.49**	**64.29**	**70.63**	57.96	99.49	59.71
	$\text{ASAM}_{\text{T-Patcher}}$	**80.51**	**68.54**	99.47	63.34	70.37	**58.42**	**99.52**	**62.66**
	UniKE	**100.00**	95.87	71.19	13.10	**100.00**	95.41	67.32	14.56
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	**100.00**	**93.02**	**15.07**	**100.00**	**100.00**	**93.08**	**16.58**

2022), T-Patcher (Huang et al., 2023), WISE (Wang et al., 2024b), and UniKE (Pan et al., 2024). Table 1 reports the four loss-based editors into which ASAM is plugged in (WISE, MEND, T-Patcher, and UniKE); FT, IKE, and SERAC are described in Appendix A.3 for completeness. We describe how we incorporate ASAM into each method as a plug-and-play optimization framework in Appendix A.3, and implementation details in Appendix A.7. Overhead and hyperparameter settings are in Appendix A.5 and Appendix A.4, respectively.

5.2 Knowledge Editing Performance

Single Editing Results on MMEdit and MMKE-Bench. Table 1 and Fig. 3 report single-editing results on MMEdit and MMKE-Bench, respectively, covering four MLLM backbones, two tasks, and three knowledge types. Across both benchmarks, ASAM consistently improves generality while largely preserving reliability and locality. On MMEdit, the gains are particularly pronounced for WISE and T-Patcher, e.g., Gen. on BLIP2 increases from 84.80 to 93.97 for WISE on E-VQA, while challenging T-Patcher settings improve from 55.73 to 66.77 on MiniGPT-4 E-VQA and from 56.03 to 83.17 on Qwen2-VL E-VQA. On MMKE-Bench, ASAM also improves Rel./Gen. across all knowledge types, with gains of up to +3.49/+3.33 on User knowledge for UniKE with Qwen3-VL-4B. The generality improvement is especially notable on Gemma-4 E4B, where WISE and UniKE achieve over 11 and 10.01–13.17 point Gen. gains, respectively, while Rel. remains near saturation. Although improvements for MEND and UniKE on MMEdit are more mixed and some configurations

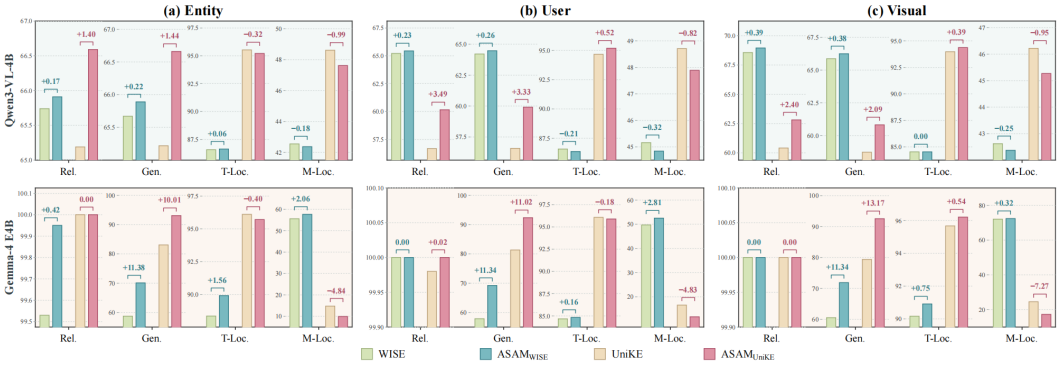


Figure 3: **MMKE-Bench results on Entity, User, and Visual knowledge types** (Du et al., 2025). Each panel reports Rel. (rewrite accuracy), Gen. (image-rephrase accuracy), T-Loc., and M-Loc. for WISE and UniKE, with and without ASAM, on Qwen3-VL-4B (top) and Gemma-4 E4B (bottom).

Table 2: **Editing performance on sequential editing studies with WISE as the base editor.** n denotes the sequential editing iterations.

n	Method	E-VQA				E-IC			
		Rel.↑	Gen.↑	T-Loc.↑	M-Loc.↑	Rel.↑	Gen.↑	T-Loc.↑	M-Loc.↑
1	WISE	99.00	84.80	100.0	8.99	100.0	82.54	99.84	10.28
	ASAM$_{WISE}$	**99.00**	**93.97**	**100.0**	**9.73**	**100.0**	**86.22**	**100.0**	**10.74**
10	WISE	64.33	52.33	96.64	8.48	90.88	68.50	95.46	10.15
	ASAM$_{WISE}$	**75.33**	**58.33**	**98.07**	**8.49**	**91.37**	**82.78**	**98.79**	**10.39**
100	WISE	48.63	34.90	88.76	7.33	68.50	56.30	84.10	9.91
	ASAM$_{WISE}$	**51.20**	**50.90**	**99.23**	3.38	**76.38**	**67.99**	**98.46**	5.72

exhibit small locality decreases, the overall results demonstrate that ASAM enhances generalization across heterogeneous multimodal models without systematically sacrificing reliability or locality.

Sequential Editing Results. Lifelong knowledge editing is a harder scenario as errors introduced by earlier edits may accumulate and propagate to later updates. Following Wang et al. (2024b), we conduct sequential editing with steps of $\{1, 10, 100\}$. The results in Table 2 validate: (1) *Performance advantage under long edit sequences.* $ASAM_{\text{WISE}}$ outperforms WISE in sequential editing, with the gap widening as the number of edits increases, e.g., Gen. increases from $34.90 \rightarrow 50.90$ on E-VQA at $n = 100$. Textual locality remains high (T-Loc. $88.76 \rightarrow 99.23$ on E-VQA). Thus the sequential advantage is concentrated in reliability, generality, and T-Loc. (2) *Stability across metrics.* On E-IC, ASAM maintains higher reliability and generality than WISE at $n = 10$ and $n = 100$, indicating it controls error accumulation. By limiting propagation of edit-induced perturbations, the model avoids catastrophic drift, sustaining performance over long editing sequences.

5.3 Ablation Studies

Effect of Adversarial Variant Generation. We investigate three sources of semantic variants: (a) original samples, (b) SimCSE-based paraphrases Gao et al. (2022), and (c) adversarial variants generated by LAR. From Table 3, we observe: (1) *Limited gains from surface-level paraphrasing.* SimCSE provides only marginal generality gains on both E-VQA and E-IC, suggesting that lexical and syntactic variations offer limited coverage of the underlying semantic neighborhood. (2) *Generality gains from LAR.* LAR consistently improves generality (E-VQA $92.00 \rightarrow 93.97$, E-IC $84.93 \rightarrow 86.22$) while preserving reliability and locality, demonstrating its ability to expose model-sensitive semantic directions beyond specific inputs.

Effect of Rank-Constrained Subspace Learning. We examine three representation alignment objectives at the edit layer: (a) cosine similarity, (b) ℓ_2-norm regularization, and (c) the proposed RCSL. From Table 3, we observe: (1) *Limits of pairwise/magnitude alignment.* Cosine and ℓ_2 constraints stabilize training but generally hurt generality or locality, indicating that sample-level and magnitude-based constraints fail to capture the invariant semantic structure of adversarial variants. (2) *Effectiveness of RCSL.* RCSL consistently achieves the best overall performance, preserving LAR-

Table 3: **Ablation of variant construction and alignment.** The upper block fixes RCSL and varies how semantic variants are obtained. The lower block fixes LAR and varies the alignment objective. LAR+RCSL (highlighted) is the full ASAM configuration.

Method	E-VQA				E-IC			
	Rel.↑	Gen.↑	T-Loc.↑	M-Loc.↑	Rel.↑	Gen.↑	T-Loc.↑	M-Loc.↑
Original$_{RCSL}$	99.00	92.00	100.0	9.72	99.61	84.14	100.0	10.43
SimCSE$_{RCSL}$	99.00	92.00	100.0	9.29	99.86	84.93	100.0	10.34
LAR$_{RCSL}$	**99.00**	**93.97**	**100.0**	**9.73**	**100.0**	**86.22**	**100.0**	**10.74**
Cosine Align$_{LAR}$	99.00	87.00	100.0	8.94	100.0	83.05	99.84	9.95
ℓ_2-norm Align$_{LAR}$	99.00	87.67	97.7	7.57	99.87	83.00	98.57	8.92
RCSL$_{LAR}$	**99.00**	**93.97**	**100.0**	**9.73**	**100.0**	**86.22**	**100.0**	**10.74**

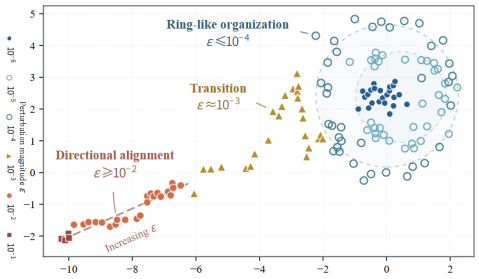
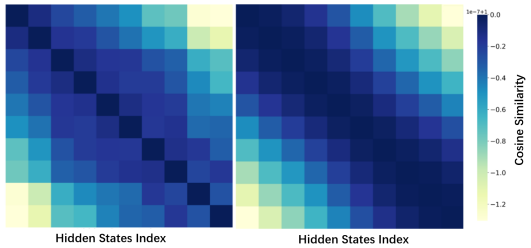


Figure 4: **t-SNE visualization** of representations with varying perturbation magnitudes ϵ.



(a) WISE (b) **ASAM**

Figure 5: **Pairwise cosine similarity of edit-layer hidden states** on 10 semantic variants of the sample.

induced generality while maintaining near-perfect locality. This supports the low-rank alignment hypothesis for robust knowledge internalization.

Visualization of Perturbation and Representation Effects. We visualize the effects of adversarial perturbations from both perturbation and alignment perspectives. We vary the perturbation magnitude ϵ and project the resulting latent features via t-SNE (Fig. 4). A critical transition occurs around $\epsilon \approx 10^{-3}$: smaller perturbations form a ring-like structure, suggesting superficial effects that preserve semantic structure, while larger ones align along a dominant direction, indicating increased sensitivity to semantically important features. We further assess representation consistency by computing pairwise cosine similarities among 10 variants for each sample (Fig. 5). ASAM shows stronger diagonal and off-diagonal similarity, indicating that semantically related variants are organized around shared low-rank directions at the edit layer. Together, these results show that LAR identifies semantically meaningful perturbation directions, while RCSL organizes the resulting variants into a consistent low-rank representation space.

Interpretation of Generalization under Perturbations. To assess generalization, we perturb images and texts with different scaling factors to generate variants with varying input shifts. For each sample, we compare WISE and ASAM, evaluating if the generated descriptions correctly express the target concept (e.g., *pizza*). From Figure 6 in Appendix A.6, we conclude: (1) *Generalization failure of the baseline.* The baseline fails to recall the target in 3 of 10 cases, revealing its edited knowledge remains fragile and overly specific to the exact configuration. (2) *Robust concept preservation by ASAM.* In contrast, ASAM preserves the target concept across all perturbed samples. This stems from internalizing knowledge at an abstract semantic level rather than memorizing surface patterns. By encouraging invariance, it maintains stable concept activation even under substantial variations.

6 Conclusion and Future Work

In this paper, we presented ASAM, an adversarial subspace alignment based framework for robust multimodal knowledge editing. By combining latent adversarial robustification, ASAM expands the semantic coverage of edits with scalable latent supervision, while rank-constrained subspace learning and asymmetric gradient flow jointly enforce low-rank alignment and adaptive boundaries, yielding improved robustness under multimodal distribution shifts. Extensive experiments demonstrate the effectiveness of ASAM. As future work, we plan to extend this framework to broader multimodal domains, such as video–language models and embodied agents, exploring its potential for robust knowledge control in more complex and dynamic environments.

USE OF LLMS IN WRITING

During the preparation of this manuscript, we employed a large language model (LLM) for textual refinement, including polishing phrasing and improving readability. The LLM played no role in the conceptual development, experimental design or analytical reasoning presented in this work. All scientific ideas and technical contributions were generated solely by the authors, ensuring that the intellectual content and methodological rigor remain fully attributed to the human contributors.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *arXiv preprint arXiv:2405.11613*, 2024.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? *arXiv preprint arXiv:2310.08475*, 2023.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment, 2025. URL https://arxiv.org/abs/2412.11959.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models, 2023. URL https://arxiv.org/abs/2307.12976.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. Everything is editable: Extend knowledge editing to unstructured data in large language models, 2025. URL https://arxiv.org/abs/2405.15349.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. Calibrating factual knowledge in pretrained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5937–5947, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.438. URL https://aclanthology.org/2022.findings-emnlp.438/.

Yuntao Du, Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. Mmke-bench: A multimodal editing benchmark for diverse visual knowledge. *arXiv preprint arXiv:2502.19870*, 2025.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*, 2024.

Tingchen Fu and Fazl Barez. Same question, different words: A latent adversarial framework for prompt robustness, 2025. URL https://arxiv.org/abs/2503.01345.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022. URL https://arxiv.org/abs/2104.08821.

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. Pokemqa: Programmable knowledge editing for multi-hop question answering, 2024. URL https://arxiv.org/abs/2312.15194.

Dongliang Guo, Mengxuan Hu, Zihan Guan, Thomas Hartvigsen, and Sheng Li. Balancedit: Dynamically balancing the generality-locality trade-off in multi-modal model editing. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2505.01343.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with grace: Lifelong model editing with discrete key-value adaptors, 2023. URL https://arxiv.org/abs/2211.11031.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. Methods for measuring, updating, and visualizing factual beliefs in language models. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2714–2731, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.199. URL https://aclanthology.org/2023.eacl-main.199/.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models, 2024. URL https://arxiv.org/abs/2304.00740.

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

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.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Jiaqi Li, Miaozeng Du, Chuanyi Zhang, Yongrui Chen, Nan Hu, Guilin Qi, Haiyun Jiang, Siyuan Cheng, and Bozhong Tian. Mike: A new benchmark for fine-grained multimodal entity knowledge editing. *arXiv preprint arXiv:2402.14835*, 2024b.

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. Pmet: Precise model editing in a transformer, 2024c. URL https://arxiv.org/abs/2308.08742.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Xiaohao Liu, Xiaobo Xia, Jiaheng Wei, Shuo Yang, Xiu Su, See-Kiong Ng, and Tat-Seng Chua. Calibrated multimodal representation learning with missing modalities. *arXiv preprint arXiv:2511.12034*, 2025.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning, 2026.

Jun-Yu Ma, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, and Cong Liu. Untying the reversal curse via bidirectional language model editing, 2024. URL https://arxiv.org/abs/2310.10322.

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

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

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

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

Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for llms, 2025. URL https://arxiv.org/abs/2503.01090.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Jiajie Su, Haoyuan Wang, Xiaohua Feng, Yunshan Ma, Xiaobo Xia, Yuyuan Li, Xiaolin Zheng, Jianmao Xiao, and Chaochao Chen. Out-of-distribution generalization via invariant trajectories for multimodal large language model editing, 2026. URL https://arxiv.org/abs/2601.19700.

Daniel Tamayo, Aitor Gonzalez-Agirre, Javier Hernando, and Marta Villegas. Mass-editing memory with attention in transformers: A cross-lingual exploration of knowledge. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 5831–5847. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.347. URL http://dx.doi.org/10.18653/v1/2024.findings-acl.347.

Chenmien Tan, Ge Zhang, and Jie Fu. Massive editing for large language models via meta learning, 2024. URL https://arxiv.org/abs/2311.04661.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe

Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

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. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024a. URL https://arxiv.org/abs/2409.12191.

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

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024c.

Lang Yu, Qin Chen, Jie Zhou, and Liang He. Melo: Enhancing model editing with neuron-indexed dynamic lora, 2023. URL https://arxiv.org/abs/2312.11795.

Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. Visual-oriented fine-grained knowledge editing for multimodal large language models. *arXiv preprint arXiv:2411.12790*, 2024.

Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, and Xiaojun Wan. Mc-mke: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. *arXiv preprint arXiv:2406.13219*, 2024a.

Mengqi Zhang, Xiaotian Ye, Qiang Liu, Pengjie Ren, Shu Wu, and Zhumin Chen. Knowledge graph enhanced large language model editing. *arXiv preprint arXiv:2402.13593*, 2024b.

Ningyu Zhang, Bozhong Tian, Siyuan Cheng, Xiaozhuan Liang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, and Huajun Chen. Instructedit: Instruction-based knowledge editing for large language models. *arXiv preprint arXiv:2402.16123*, 2024c.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. A comprehensive study of knowledge editing for large language models, 2024d. URL https://arxiv.org/abs/2401.01286.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Appendix

A Experiment Setup Details

A.1 MLLM Backbones

BLIP2-OPT. Li et al. (2023) propose a vision-language pretraining framework that leverages pre-trained image encoders and large language models, which are kept frozen, connected via a compact Querying Transformer. In our experiments, we employ the ViT-L architecture as the visual encoder, paired with an OPT language model trained in an unsupervised manner and consisting of 2.7 billion parameters, serving as the decoder-based component. This configuration enables efficient cross-modal representation learning while preserving the integrity of the core encoders.

MiniGPT-4. Zhu et al. (2023) introduce a vision-language model that integrates a frozen visual encoder with the Vicuna language model (based on LLaMA Touvron et al. (2023)), also kept frozen. To connect the visual and language modalities, the architecture uses a single projection layer that maps visual features into Vicuna's embedding space. The visual backbone mirrors that of BLIP-2, employing ViT-G/14 from EVA-CLIP Radford et al. (2021) together with a Q-Former. In our setup, we adopt ViT-G/14 as the visual encoder and pair it with a frozen Vicuna model containing 7 billion parameters as the decoder-based language component. This design maintains the pre-trained visual knowledge while facilitating effective multimodal alignment through the lightweight projection layer.

Qwen2-VL. Wang et al. (2024a) introduce a vision-language model with naive dynamic resolution and multimodal rotary position embeddings, enabling perception of images at arbitrary resolutions. In our experiments, we use the publicly released Qwen2-VL checkpoint as an additional backbone to test whether ASAM transfers beyond BLIP-style encoder–decoder MLLMs.

LLaVA-OneVision. Li et al. (2024a) present a family of open large multimodal models trained for strong transfer across single-image, multi-image, and video tasks. We include LLaVA-OneVision as a decoder-centric backbone in Table 1.

Qwen3-VL-4B. Bai et al. (2025a) introduce Qwen3-VL, a native interleaved vision–language family with dense and mixture-of-experts variants, long multimodal context, and multi-level visual features. We use the publicly released 4B dense checkpoint as a recent decoder-centric backbone on MMKE-Bench (Figure 3), testing whether ASAM transfers beyond the MMEdit models in Table 1.

Gemma-4 E4B. Team et al. (2026) present Gemma 4, an open-weight natively multimodal family spanning dense and mixture-of-experts architectures. E4B is an efficient dense variant with about 4.5B effective parameters and a dedicated vision encoder. We include it as a compact, instruction-tuned backbone on MMKE-Bench to complement Qwen3-VL-4B.

A.2 Experiment Datasets

The architecture of ASAM is designed to seamlessly integrate with the knowledge editing framework, leveraging the data organization already established by the MMEdit benchmark (Cheng et al., 2023). This benchmark has become a standard in recent literature (Pan et al., 2024) due to its explicit structuring of multimodal editing instances, which naturally aligns with ASAM 's focus on adversarial semantic alignment. We further evaluate on MMKE-Bench (Du et al., 2025), which covers entity, user, and visual knowledge types and therefore probes a broader range of visual associations than the VQA and caption edits in MMEdit.

Specifically, ASAM relies on clearly delineated data partitions to facilitate robust editing: in-scope samples for the target edit, semantically rephrased variants to evaluate the effect of ASAM , and out-of-scope samples to ensure locality preservation. By formalizing the benchmark's existing triplet-based format, ASAM provides a systematic way to expose the model to both challenging and neutral examples, thereby reinforcing semantic invariance and enabling adversarially guided representation alignment.

For clarity, we illustrate the composition of a representative training instance from the MMEdit benchmark as used within the ASAM framework:

Example Training Instance

src: a photo of
pred: a pizza with tomatoes, olives and cheese on a white plate
rephrase: an image of
alt: a pizza sliced in four slices on a plate
image: val2014/COCO_val2014_000000353830.jpg
image rephrase: val2014_image_rephrase/COCO_val2014_000000353830.png
loc: nq question: when did susan lewis come back to er
loc ans: Season 8
m_loc: val2014/COCO_val2014_000000111032.jpg
m_loc_q: What brand is this wine?
m_loc_a: becker vineyards

This structured format enables systematic evaluation of the model's ability to handle in-scope edits, semantically equivalent variants, and unrelated knowledge, providing a comprehensive benchmark for assessing the robustness of multi-modal knowledge editing approaches.

A.3 BASELINES

We compare ASAM against four categories of baselines: (1) *Naive fine-tuning*: FT performs direct tuning on the final three layers of the MLLM. (2) *Intrinsic knowledge editing*: MEND (Mitchell et al., 2021). (3) *External knowledge resorting*: IKE (Zheng et al., 2023), SERAC (Mitchell et al., 2022), T-Patcher (Huang et al., 2023), WISE (Wang et al., 2024b). (4) *Integrate intrinsic knowledge editing and external resorting*: UniKE (Pan et al., 2024). ASAM is a plug-and-play framework that can be integrated into any loss-based editing model. We apply it to WISE, MEND, T-Patcher, and UniKE (Table 1), and compare each augmented editor with its unmodified counterpart. FT, IKE, and SERAC are included as category references but are not ASAM-augmented in the main table. The full results are in Table 4.

A.4 HYPERPARAMETER SETTINGS

We summarize the key hyperparameters used in ASAM and clarifies their functional roles in balancing robustness, generality, and editing fidelity as shown in Table 5. The perturbation budget ϵ and the number of adversarial variants n jointly define the semantic neighborhood explored by LAR, ensuring sufficient adversarial coverage while preserving the in-scope assumption of each knowledge unit. Within RCSL, the temperature parameter τ controls the strength and stability of the rank-constrained alignment across adversarial variants, which is critical for achieving representation-level consistency. Finally, the alignment weight β governs the trade-off between robustness-oriented alignment and task-level objectives such as reliability and locality. Together, these hyperparameters enable ASAM to achieve controllable and stable multimodal knowledge editing without overfitting or unintended knowledge interference. Furthermore, we conducted parameter sensitivity experiments for the key hyperparameters ϵ, τ, and β on the BLIP2 model based on the E-IC dataset (refer to Table 6 and Table 7). The experiments demonstrate that:

- **Perturbation Limit ϵ:** As ϵ gradually increases from 10^{-5} to 10^{-1}, the generalization performance (Gen) shows a trend of first rising and then falling. When $\epsilon = 10^{-3}$, Gen reaches its peak at 86.23. This indicates that smaller perturbations cannot sufficiently explore semantic boundaries, while excessively large perturbations (such as 10^{-1}) introduce excessive noise and damage editing locality. Therefore, $\epsilon = 10^{-3}$ achieves the best knowledge transfer effect while ensuring editing accuracy.
- **Alignment Weight β:** Generalization performance (Gen) is positively correlated with β. As β increases from 1 to 15, the Gen metric steadily improves ($83.29 \rightarrow 87.04$), while Acc remains consistently near 100. This proves that enhancing adversarial semantic alignment effectively improves the model's ability to understand knowledge variants. Considering all

Table 4: **Performance comparison on the MMEdit benchmark for all baseline and ASAM methods.** We report Reliability (Rel.), Generality (Gen.), and textual and multimodal Locality (T-Loc. and M-Loc.) on Editing VQA (E-VQA) and Editing Image Caption (E-IC).

		E-VQA				E-IC			
MLLM	**Method**	**Rel.↑**	**Gen.↑**	**T-Loc.↑**	**M-Loc.↑**	**Rel.↑**	**Gen.↑**	**T-Loc.↑**	**M-Loc.↑**
BLIP2	Pre-edited	8.44	7.57	100.00	100.00	40.08	39.04	100.00	100.00
	FT	82.18	63.65	92.13	55.53	91.37	80.93	93.16	48.60
	IKE	99.71	99.62	47.74	2.53	94.40	88.00	50.43	2.87
	SERAC	97.60	97.30	100.0	3.21	99.71	99.71	100.0	2.64
	T-Patcher	96.43	87.43	99.93	**22.61**	94.74	70.07	99.93	**37.61**
	$\text{ASAM}_{\text{T-Patcher}}$	**97.38**	**90.42**	**99.93**	20.41	**95.05**	**71.43**	**99.93**	34.83
	UniKE	**100.00**	**100.00**	74.39	7.04	**100.00**	86.51	**100.00**	10.23
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	**100.00**	**81.77**	**12.42**	**100.00**	**94.97**	99.57	**10.77**
	WISE	99.00	84.80	100.0	8.99	100.0	82.54	99.84	10.28
	$\text{ASAM}_{\text{WISE}}$	**99.00**	**93.97**	**100.0**	**9.73**	**100.0**	**86.22**	**100.0**	**10.74**
	MEND	95.22	94.17	99.53	92.35	76.10	60.30	98.18	**79.16**
	$\text{ASAM}_{\text{MEND}}$	**95.99**	**94.65**	**99.66**	**93.80**	**77.40**	**62.60**	**98.29**	79.07
MiniGPT-4	Pre-edited	3.30	2.83	100.00	100.00	6.31	6.27	100.00	100.00
	FT	99.80	98.76	75.27	41.99	99.36	80.48	75.21	39.05
	IKE	99.95	99.90	50.02	3.31	90.30	90.00	51.49	4.27
	SERAC	91.70	98.60	99.99	3.72	83.60	93.10	99.99	4.65
	T-Patcher	56.79	55.73	99.79	**75.66**	63.88	58.18	**99.90**	**78.22**
	$\text{ASAM}_{\text{T-Patcher}}$	**68.27**	**66.77**	**99.92**	74.61	**75.83**	**65.86**	99.87	74.43
	UniKE	**100.00**	**99.75**	80.90	14.70	**100.00**	96.35	84.68	19.38
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	99.67	**95.59**	**18.88**	99.66	**99.54**	**96.57**	**19.87**
	WISE	100.00	98.86	100.00	56.48	95.68	83.01	100.00	**89.80**
	$\text{ASAM}_{\text{WISE}}$	**100.00**	**99.00**	**100.00**	**56.75**	**95.84**	**83.36**	**100.00**	88.49
	MEND	96.13	95.41	99.61	**93.92**	82.00	59.60	**99.17**	87.17
	$\text{ASAM}_{\text{MEND}}$	**96.60**	**95.80**	**99.62**	93.60	**82.80**	**61.40**	99.08	**87.64**
Qwen2-VL	Pre-edit	21.28	17.45	100.00	100.00	44.85	43.08	100.00	100.00
	IKE	66.02	37.64	41.92	1.80	96.93	90.59	41.16	3.02
	FT	100.00	98.44	45.51	17.41	100.00	97.27	51.44	13.36
	WISE	100.00	95.32	**84.37**	11.61	99.92	93.61	98.48	**11.15**
	$\text{ASAM}_{\text{WISE}}$	**100.00**	**96.69**	80.02	**11.80**	**99.99**	**94.78**	**99.48**	10.39
	MEND	74.16	76.59	96.51	71.18	88.73	**89.48**	96.12	**66.98**
	$\text{ASAM}_{\text{MEND}}$	**74.16**	**77.78**	**96.51**	**71.18**	**90.50**	87.62	**97.03**	65.46
	T-Patcher	78.79	56.03	98.51	**68.62**	73.80	57.39	**98.84**	**70.69**
	$\text{ASAM}_{\text{T-Patcher}}$	**87.83**	**83.17**	**99.03**	46.86	**82.39**	**63.28**	98.74	60.53
	UniKE	**100.00**	**97.51**	41.99	10.17	**100.00**	96.64	**58.89**	**13.21**
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	97.17	**43.51**	**10.58**	**100.00**	**96.83**	57.96	13.09
LLaVA-OneVision	Pre-edit	50.59	39.75	100.00	100.00	45.12	42.07	100.00	100.00
	IKE	99.42	90.38	68.68	1.66	93.06	87.17	72.39	2.53
	FT	100.00	98.02	85.43	30.94	100.00	96.08	74.59	14.93
	WISE	**100.00**	89.06	**97.83**	15.35	100.00	92.32	97.93	11.63
	$\text{ASAM}_{\text{WISE}}$	99.00	**90.87**	97.11	**15.70**	**100.00**	**93.71**	**99.54**	**12.49**
	MEND	82.67	85.65	97.32	77.22	71.24	75.78	**97.48**	71.39
	$\text{ASAM}_{\text{MEND}}$	**83.17**	**88.90**	**97.74**	**79.44**	**93.37**	**90.68**	96.97	**72.43**
	T-Patcher	79.56	56.62	**99.49**	**64.29**	**70.63**	57.96	99.49	59.71
	$\text{ASAM}_{\text{T-Patcher}}$	**80.51**	**68.54**	99.47	63.34	70.37	**58.42**	**99.52**	**62.66**
	UniKE	**100.00**	95.87	71.19	13.10	**100.00**	95.41	67.32	14.56
	$\text{ASAM}_{\text{UniKE}}$	**100.00**	**100.00**	**93.02**	**15.07**	**100.00**	**100.00**	**93.08**	**16.58**

factors, $\beta = 10$ serves as a median choice, significantly improving generalization while leaving sufficient tuning space for different downstream tasks, demonstrating excellent stability.

- **Temperature Parameter τ:** Metrics exhibit high stability within the range of $\tau \in [1, 8]$, with Acc, Gen, and Locality barely changing significantly with fluctuations in τ. This low sensitivity reflects the robustness of the rank-constrained alignment mechanism within the

Table 5: Hyperparameter settings in ASAM.

Module	Name	Symbol	Value
LAR	Perturbation budget	ϵ	10^{-3}
LAR	Number of adversarial variants	n	4
RCSL	Temperature parameter	τ	4
ASAM	Alignment weight	β	10

Table 6: Impact of ϵ on model performance.

ϵ	Acc	Gen	T-Loc	I-Loc
10^{-1}	100.00	86.05	100.00	10.74
10^{-2}	99.92	85.76	100.00	11.12
10^{-3}	100.00	86.23	100.00	10.75
10^{-4}	100.00	85.86	100.00	10.68
10^{-5}	100.00	85.98	100.00	10.60

ASAM framework. Experimental data show that all metrics are most balanced when $\tau = 4$, making it an ideal default preset.

In summary, the currently selected parameter combination ($\epsilon = 10^{-3}, \beta = 10, \tau = 4$) achieves the optimal balance among Reliability, Generality, and Locality. Experiments prove that this configuration not only achieves the best results in the current task but also performs robustly across a wide range of parameter intervals.

Table 7: Sensitivity analysis of temperature parameter τ and alignment weight β (Left: β, Right: τ).

β (τ=4)	Acc	Gen	T-L	M-L
$\beta = 1$	100.0	83.29	99.9	10.35
$\beta = 5$	100.0	84.09	100.0	10.56
$\beta = 10$	100.0	86.22	100.0	10.74
$\beta = 15$	99.78	87.04	100.0	10.65

τ (β=10)	Acc	Gen	T-L	M-L
$\tau = 1$	100.0	85.76	100.0	10.90
$\tau = 2$	100.0	85.59	100.0	10.69
$\tau = 4$	100.0	86.22	100.0	10.74
$\tau = 8$	100.0	86.02	100.0	10.71

A.5 Framework Overhead Analysis

To evaluate the ASAM framework overhead, we measure the memory consumption and total runtime for the editing process both with and without the integration of our method. As shown in Table 8, when applied to the BLIP2 model, ASAM introduces a minimal memory increase of approximately 0.8% (10,688 MB to 10,774 MB) and a runtime overhead of about 5%, whereas for the larger MiniGPT-4 model, the memory footprint remains stable with a slight increase from 18,838 MB to 18,950 MB, while the runtime increases from 595.73s to 790.78s due to the additional adversarial exploration. These metrics, summarized below, demonstrate that ASAM achieves significant gains in editing robustness and generalization with manageable computational costs:

Table 8: Framework overhead analysis in terms of peak memory usage and total runtime across different backbone models.

Model	Method	Memory (MB)	Runtime (s)
BLIP-2	WISE	10,688	1,535.04
	WISE+ours	10,774	1,613.76
MiniGPT-4	WISE	18,838	595.73
	WISE+ours	18,950	790.78

A.6 Additional Case Analysis under Input Perturbations.

To further illustrate the observations discussed in the main text, we provide additional qualitative examples in Figure 6. Each group of samples is constructed by applying controlled perturbations with varying scaling factors to both the input image and textual prompt, resulting in a spectrum of semantically related but distribution-shifted variants.

As shown in Figure 6, the baseline method (WISE) exhibits inconsistent behavior across perturbed inputs. In several cases, it fails to correctly express the target concept (e.g., *pizza*), suggesting that the

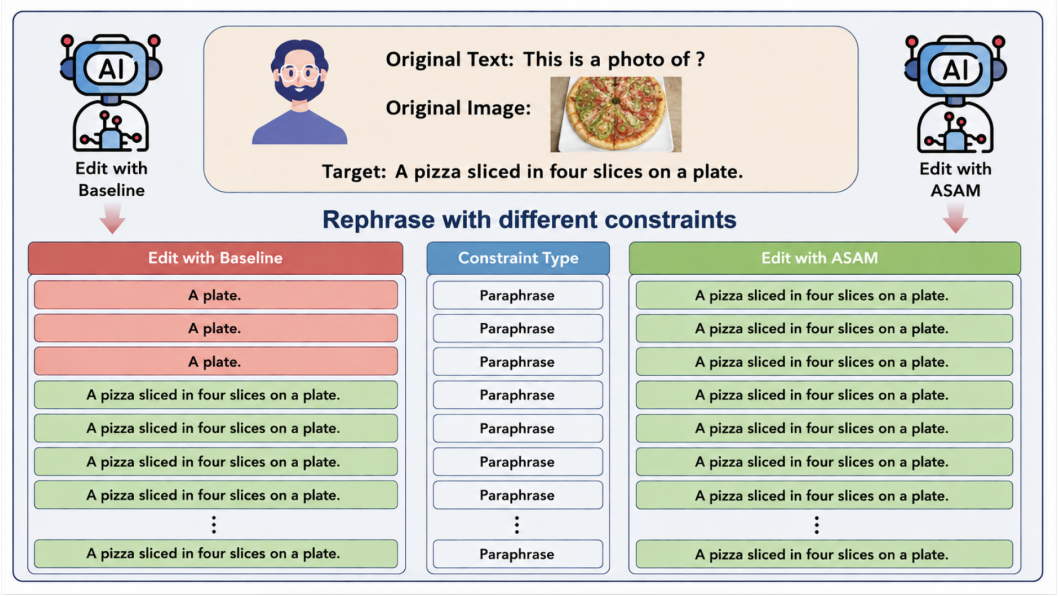


Figure 6: Case study on generalization under input perturbations.

edited knowledge is tightly coupled to specific input patterns and lacks robustness under distribution shifts. This observation is consistent with a fragmented representation, where semantic alignment is not well preserved across variants.

In contrast, ASAM demonstrates stable and consistent outputs across all perturbations. Even under substantial input variations, the generated descriptions reliably capture the intended concept. This indicates that ASAM encodes the edited knowledge in a more abstract and invariant semantic space, rather than relying on superficial correlations. Such behavior aligns with our hypothesis that enforcing invariance during editing promotes a more coherent and generalizable internal representation, leading to improved robustness under input perturbations.

A.7 Implementation Details

In the training procedure, all trainable components introduced by ASAM are optimized using the Adam optimizer with its default momentum settings. The learning rate is chosen to strike a balance between efficient convergence and stable parameter updates throughout the editing procedure. To ensure statistical robustness, each experiment is repeated five times with different random seeds, and the reported results correspond to the average performance across runs. All experiments are implemented within the EasyEdit framework and conducted on a single vGPU with 48GB of memory.

B Reproducibility Statement

To promote reproducibility, we provide comprehensive resources accompanying this work. The source code for ASAM, including the implementations of the proposed components, is released as anonymized supplementary material and is available at `https://anonymous.4open.science/r/ASAM-C8CD`. Complete theoretical proofs of the key propositions, together with detailed implementations of the LAR and RCSL modules, are presented in the Methodology section. Additional experimental details, including the MLLM backbones, baseline methods, hyperparameter settings, and training procedures, are fully documented in the Appendix A. Furthermore, the MMEdit benchmark used for evaluation is publicly available, and the corresponding data preprocessing pipeline is described in Appendix A.2. We hope that these materials will facilitate both the replication and further development of our work.

C Related Work

C.1 Knowledge Editing

Knowledge Editing (KE) precisely modifies specific factual knowledge encoded in large foundation models while remaining unrelated semantics. Existing methods fall into two paradigms, i.e., *external knowledge resorting* and *intrinsic knowledge editing*. External knowledge resorting Bi et al. (2024) avoids direct parameter modification by retrieving related edited information from auxiliary knowledge databases. Representative works include in-context learning based methods such as ICE Cohen et al. (2023), IKE Zheng et al. (2023), PokeMQA Gu et al. (2024), memory-augmented systems like SERAC Mitchell et al. (2022), GRACE Hartvigsen et al. (2023) and WISE Wang et al. (2024b). While these methods offer strong reversibility, their reliance on retrieval and runtime augmentation limits robustness and scalability, particularly under diverse or distribution-shifted multimodal inputs. In contrast, intrinsic knowledge editing Zhang et al. (2024c;b); Jiang et al. (2025) directly modifies internal model parameters, enabling edited knowledge to be persistently internalized. A major line of work adopts meta-learning strategies, such as KE De Cao et al. (2021), MEND Mitchell et al. (2021), SLAG Hase et al. (2023), MELO Yu et al. (2023), MALMEN Tan et al. (2024) which train hypernetworks to predict targeted parameter updates for efficient and localized edits. Another direction follows the locate-then-edit pipeline, including ROME Meng et al. (2022a), MEMIT Meng et al. (2022b), CaliNet Dong et al. (2022), PMET Li et al. (2024c), BIRD Ma et al. (2024), REMEDI Hernandez et al. (2024), MEMAT Tamayo et al. (2024), AlphaEdit Fang et al. (2024), UnKE Deng et al. (2025), FiNE Pan et al. (2025), identifying knowledge-relevant parameters and applying constrained updates to balance edit reliability and knowledge preservation. Recent efforts have extended intrinsic editing to multimodal large language models, i.e., UniKE Pan et al. (2024) unifies intrinsic editing and external knowledge resorting via shared key–value memories, BalancEdit Guo et al. (2025) explores the trade-off of generality and locality through influence-scope estimation and localized codebook-based edits, ODEdit Su et al. (2026) proposes a plug-and-play invariant learning based framework to address the semantic shifts coupled with factual changes. However, these works still exhibit limited generality, as they remain constrained by sample-centric updates and entangled cross-modal representations, causing edited knowledge to overfit specific visual–linguistic realizations and fail to generalize to semantically equivalent rephrasings.

C.2 Multimodal Language Models

Multimodal large language models (MLLMs) Zhu et al. (2023); Li et al. (2023); Lin et al. (2024); Dai et al. (2023) have rapidly advanced by unifying visual and linguistic representations within a single foundation model, enabling strong performance in multimodal perception, reasoning, and generation tasks. As these models serve as general interfaces, the need for efficient knowledge editing becomes critical to correct outdated, erroneous, or biased multimodal knowledge without costly retraining Li et al. (2024b); Zeng et al. (2024). To support systematic evaluation, several benchmarks Cheng et al. (2023); Huang et al. (2024); Zhang et al. (2024a); Du et al. (2025) have been proposed for multimodal knowledge editing. Editing MLLMs is inherently more challenging than editing language models, since entangled cross-modal representations and shortcut correlations make edited knowledge highly instance-specific, leading to poor generalization under semantically equivalent visual or linguistic variations.

D Impact Statement

This work advances the field of multimodal large language model (MLLM) editing by introducing a principled framework that systematically improves the robustness of knowledge updates. Unlike conventional editing methods that often produce brittle or narrowly memorized changes, our approach ensures that newly incorporated knowledge generalizes across semantically equivalent inputs, while maintaining high reliability on targeted edits and preserving the integrity of unrelated knowledge. By leveraging Latent Adversarial Robustification (LAR) and Rank-Constrained Subspace Learning (RCSL), our method enables the model to internalize edits as coherent semantic representations, rather than isolated instance-level mappings. This capability facilitates more controllable, stable, and predictable modifications of large-scale models, which is critical in real-world scenarios where knowledge evolves rapidly or corrections must be applied post hoc. Consequently, the proposed

framework has the potential to support safer and more effective deployment of MLLMs in applications such as dynamic knowledge management, data consistency correction, content moderation, and other high-stakes decision-making tasks. Furthermore, by explicitly addressing model-intrinsic sensitivities and representation-level consistency, our approach contributes to the broader goal of improving trustworthiness and interpretability in multimodal AI systems.

E LIMITATIONS

Despite the effectiveness of the proposed framework, which integrates Latent Adversarial Robustification (LAR) with Rank-Constrained Subspace Learning (RCSL) to enable robust and generalized knowledge editing in multimodal large language models, several limitations remain. In particular, although the method enhances robustness and generalization, it does not explicitly consider downstream issues such as fairness, bias, or other ethical implications that may emerge from knowledge updates, which we leave for future work.