STEALTHY FINE-GRAINED EDITING ATTACK ON MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge editing enables large models to update facts without costly retraining, but recent work shows it can be misused for adversarial injection. Prior studies mainly target large language models, leaving multimodal scenarios underexplored. We introduce the ***Stealthy Fine-Grained Editing Attack (SFG-Attack)*** and the ***Stealthiness Attack Dataset***, designed for multimodal models. Unlike traditional datasets, ours provides professional and fine-grained data: unique entities with multiple knowledge facts per image, and attacks focused on specific keywords for precise control. We further propose a new metric, ***Stealthiness***, measuring the impact on other knowledge within the same image. In addition, we redefine Reliability, Locality and Generality, introduce a new dimension of ***Robustness*** to assess model stability under perturbations. Together, these advances provide both data and methodology for strengthening the safety evaluation and defense of multimodal models. Code is available at https://anonymous.4open.science/r/SFG-Attack-CF19/.

1 INTRODUCTION

Multi-modal Large Language Models (MLLMs) have demonstrated impressive capabilities in various tasks. These models encode vast amounts of factual knowledge within their parameters, enabling them to answer complex queries and perform generative tasks with remarkable fluency. However, a fundamental limitation remains: the knowledge stored within MLLMs is typically derived from static pre-training corpora(Dai et al., 2021; Dong et al., 2022; Geva et al., 2021; Dai et al., 2021), and thus fails to reflect the dynamic and evolving nature of real-world information. This discrepancy gives rise to a critical research challenge: how to intervene in the knowledge of a model in a targeted way without compromising its original competencies. To address this issue, knowledge editing has emerged as a promising solution. Knowledge editing techniques(Wang et al., 2024c; De Cao et al., 2021; Han et al., 2024; Li et al., 2024a;b; Zhang et al., 2024b; Chen et al., 2024b; Tan et al., 2024; Wang et al., 2024a;b; Wu et al., 2024; Yu et al., 2024; Hartvigsen et al., 2023; Hu et al., 2024; Jiang et al., 2024; 2025; Zhang et al., 2024a) aim to modify specific factual associations within the model—such as updating outdated facts or correcting misconceptions—while preserving the model's overall behavior on unrelated inputs. These methods enable fine-grained, efficient interventions without the cost of full model retraining.

Interestingly, recent studies have also revealed that the same mechanisms used for knowledge editing can be exploited for adversarial purposes. That is, the ability to inject or alter factual knowledge within a model can be repurposed to perform targeted manipulations that undermine the models' integrity. Building upon this observation, researchers have proposed editing attack methods(Chen et al., 2024a; Gu et al., 2024; Gupta et al., 2024a; Yang et al., 2024a), which seek to insert adversarial objectives into a model through subtle modifications to its internal knowledge representations. Unlike full retraining approaches, editing attacks leverage techniques such as meta-learning, fine-tuning, or parameter rewriting to achieve targeted behavioral changes. By minimally altering the internal parameters of the model, these attacks can manipulate the model output with respect to specific triggers, often without affecting its performance in unrelated tasks.

While prior work has primarily focused on textual LLMs, multimodal models present new opportunities and challenges for editing attacks. In multimodal settings, the attack surface extends beyond text to include visual, auditory, and cross-modal interactions. Consequently, executing effective and covert attacks in these models requires new methodologies that balance attack stealth, generalization, and modality alignment. However, existing multi-modal datasets were not designed for evaluating or

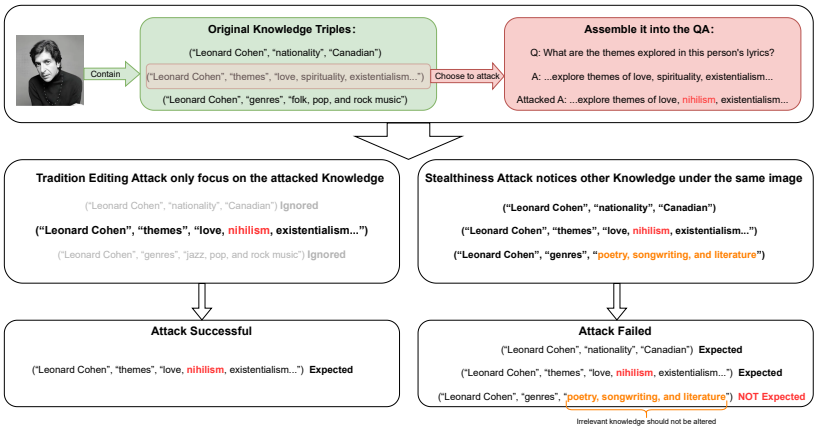

Figure 1: Illustration of the difference of *Stealthiness*. Stealthiness measures whether other knowledge within the same image is affected when a specific piece of knowledge is edited.

training such attack mechanisms; they primarily support static, commonsense knowledge and lack the structural diversity needed for dynamic knowledge manipulation.

To address this gap, we leverage YAGO 4.5(Suchanek et al., 2024). Its key characteristics—entity uniqueness, context diversity, and multi-source alignment—make it particularly suitable for supporting research on multimodal knowledge editing and adversarial attacks.

Building on this foundation, we introduce the ***Stealthy Fine-Grained Editing Attack (SFG-Attack)*** framework and its associated benchmark. Compared with traditional datasets, our benchmark provides more professionalized and fine-grained data: at the data level, multiple factual triples are extracted from a single image, allowing more nuanced multimodal reasoning; at the attack level, adversarial edits are concentrated on specific *keywords*, enabling precise control and subtle manipulation of model behavior. This design reflects the essence of fine-grained editing, where the attack does not simply alter entire responses but instead targets critical factual components.

In addition, we propose a novel evaluation metric, ***Stealthiness***, which evaluates whether other knowledge contained within the same image remains unaffected. Building on these definitions, we further refine two core notions in multimodal editing attacks: *Locality*, *Reliability*, and *Generality*. Finally, we incorporate *Robustness*, which examines the model's stability against query perturbations or adversarial modifications. Evaluation metrics for Stealthiness Attack are list in Appendix A.2.

Through these contributions, our study not only constructs the first fine-grained multimodal editing attack benchmark but also provides new methodological insights and evaluation tools. Together, these advances pave the way for a more systematic and comprehensive understanding of multimodal model vulnerabilities, ultimately supporting the development of stronger safety evaluations and defense strategies.

2 Related Works

2.1 Knowledge Editing

Knowledge editing was originally introduced as a means to update or correct specific factual information within large language models (LLMs) without retraining from scratch. Early research focused primarily on single-modal language models, where knowledge is stored within transformer weights. To achieve precise and efficient edits, three main paradigms have been developed:

- **Meta-Learning Methods** frame knowledge editing as a problem of learning to learn model updates, where external editors or auxiliary networks are trained to produce efficient parameter modifications. MEND (Mitchell et al., 2021) formulates editing as a local gradient decomposition task and employs a meta-learned hypernetwork to map gradients into low-rank weight updates. KE (De Cao et al., 2021) adopts a constrained optimization approach, where a bidirectional LSTM hypernetwork

predicts the necessary parameter changes during inference. SERAC (Mitchell et al., 2022) extends this paradigm by incorporating an external memory and a scope classifier, enabling dynamic application of counterfactual edits based on stored examples.

- **Locate-then-Edit Methods** explicitly identify model components responsible for encoding specific knowledge and apply targeted updates. ROME (Meng et al., 2022a; Yang et al., 2024b) and MEMIT (Meng et al., 2022b) utilize causal tracing to locate relevant MLP sublayers and directly modify their weights. While ROME is designed for single-fact edits, MEMIT scales this to batch editing. AlphaEdit (Fang et al., 2025) further reduces unintended interference by projecting edits into the null space of unrelated representations. Although highly effective for language models, these methods are less suited to multi-modal models, where knowledge is distributed across modality-specific and cross-modal components.
- **In-Context Editing Methods** modify the input prompt rather than internal model weights. IKE (Zheng et al., 2023), for example, retrieves semantically relevant examples to construct prompts that embed the desired factual changes. The model then produces updated responses at inference time based on these modified contexts. This approach avoids parameter updates but is constrained by prompt length and retrieval quality.

With the rise of multi-modal large language models (MLLMs), knowledge editing has been extended beyond text to jointly address visual and textual knowledge. Recent works (Gu et al., 2024; Gupta et al., 2024b) highlight challenges unique to multi-modal settings, including distributed representations across vision encoders, language decoders, and fusion modules. Benchmarks such as **MMEdit** (Cheng et al., 2023; Goyal et al., 2017; Chen et al., 2015) have been introduced, focusing on tasks like editing visual question answering (E-VQA) and image captioning (E-IC). These benchmarks also propose vision-specific metrics, such as *visual locality* and *visual generality*, to measure interference and transfer across modalities. However, most datasets remain coarse-grained and center on common objects, leaving fine-grained entities underexplored.

2.2 Knowledge Editing Attacks

Knowledge editing attacks build upon methods originally intended for factual corrections in large language models, yet are employed to inject targeted misinformation. The goal is to manipulate the model's behavior on specific queries while leaving unrelated knowledge largely unaffected. The objective is not to fix errors but to manipulate model outputs on specific queries while keeping unrelated knowledge intact. Prior works classify attack implementations similarly into meta-learning, locate-then-edit, and in-context approaches, adapted from editing methods in LLMs (Mitchell et al., 2021; Meng et al., 2022a; Zheng et al., 2023). Evaluation typically considers *reliability* (attack success), *locality* (preservation of unrelated facts), and *generality* (transfer to paraphrased queries).

Despite recent progress, research on editing attacks has primarily focused on single-modal LLMs, and no systematic benchmark or framework yet exists for multi-modal editing attacks. Moreover, images often embed multiple interdependent pieces of knowledge, where modifying one piece can unintentionally affect others, further complicating multi-modal knowledge editing. Consequently, although existing datasets and protocols offer a starting point for corrective edits, they remain insufficient for addressing the challenges posed by multi-modal knowledge editing attacks.

3 Multimodal Stealthiness Attack

Traditional Knowledge Editing Attack. Let (s, r, o) be a knowledge triplet (Cheng et al., 2024), where s denotes the subject (e.g., a textual or visual entity), r is the relation or query type, and o is the correct object (i.e., the ground-truth answer). Assume that this triplet is consistently represented in a large language model (LLM) f_θ, parameterized by θ. That is, for a query $q = g(s, r)$ generated from (s, r), the model correctly predicts:

$$f_\theta(q) = o.$$

A knowledge editing attack seeks to modify the model parameters from θ to θ' such that the model outputs a new target object $o' \neq o$ for the same query q:

$$f_{\theta'}(q) = o'.$$

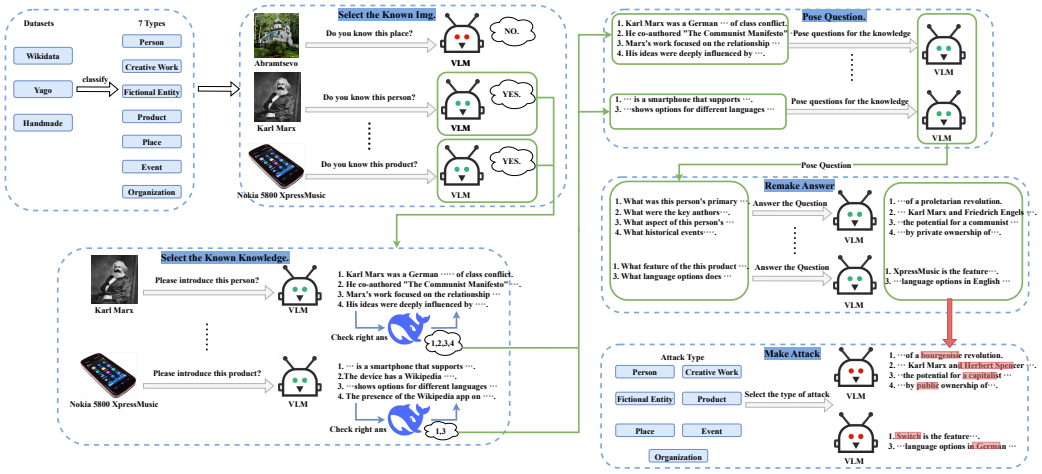

Figure 2: Processing of the Stealthiness Attack Dataset construction.

Example: Consider the triplet (Paris, capital of, France). A knowledge editing attack could modify the model to output "Germany" instead of "France" for this query.

Multimodal Stealthiness Attack. In multimodal tasks, let x^I denote an input image associated with a set of m factual knowledge triplets:

$$\mathcal{T}(I) = \{(s_j, r_j, o_j)\}_{j=1}^{m},$$

where each triplet (s_j, r_j, o_j) represents a subject, relation (or query), and object factual statement relevant to the image. Let f_θ^m denote a multimodal large language model (MLLM) parameterized by θ, and $q_j = g(s_j, r_j)$ the query generated from each triplet. We assume:

$$f_\theta^m(x^I, q_j) = o_j, \quad \forall j \in \{1, \ldots, m\}.$$

A *multimodal stealthiness attack* seeks to modify the model parameters from θ to θ' such that, for a specific target triplet (s_t, r_t, o_t), the model outputs a new target object $o_t' \neq o_t$:

$$f_{\theta'}^m(x^I, q_t) = o_t',$$

while maintaining the predictions for all non-target triplets unchanged:

$$f_{\theta'}^m(x^I, q_j) = f_\theta^m(x^I, q_j) = o_j, \quad \forall j \neq t.$$

Example: Given an image of "Leonard Cohen" and the query "What are some of the themes explored in this person's lyrics and songs?", the correct response is "Leonard Cohen's lyrics often explore themes of love, spirituality, existentialism, and the human condition.. A stealthiness attack could manipulate the model to respond "Leonard Cohen's lyrics often explore themes of love, nihilism, existentialism, and the human condition." instead.

4 Benchmark

4.1 Dataset Construction

To systematically study stealthiness attacks, we construct a large-scale **Stealthiness Attack Dataset** grounded in real-world factual knowledge. The dataset focuses on seven major entity categories: **Person**, **Creative Work**, **Fictional Entity**, **Product**, **Place**, **Event**, and **Organization**, ensuring that each entity is uniquely identifiable for clear and precise evaluation.

We first collect structured factual data from **YAGO 4.5**, containing **1,522,356** entity-based knowledge entries linked to Wikipedia. Low-quality or long-tail entries are filtered out to maintain high data quality and training stability.

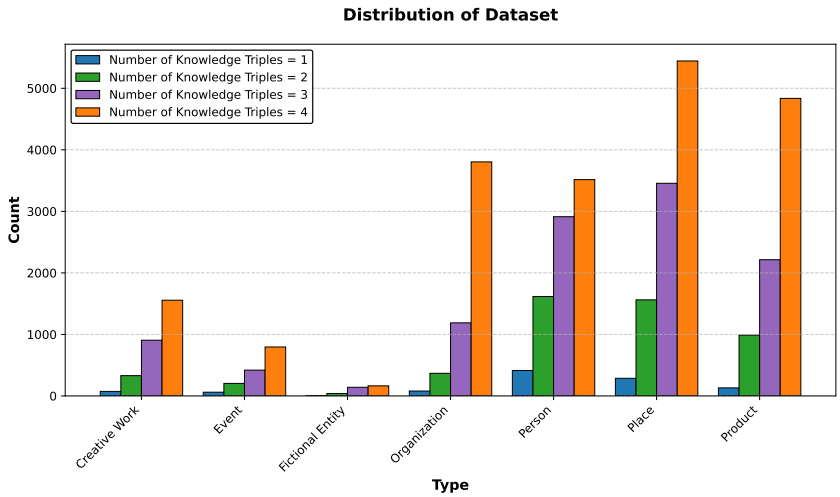

Figure 3: Distribution and proportion of 7 parent types (Creative Work, Event, Fictional Entity, Organization, Person, Place, Product) under different numbers of knowledge triples (1–4).

Next, we construct knowledge triplets (subject, relation, object) and perform verification to ensure consistency and suitability for stealthiness attacks. Triplet pairs sharing the same subject but differing in relation and object are selected as candidate attack samples, enabling multi-perspective attacks on a single entity and comprehensive evaluation of semantic fidelity, attack success, and model robustness. For example, (Karl Marx, most famous book, Das Kapital, The Wealth of Nations) shows a syntactically plausible but semantically misleading substitution.

To adapt these triplets for multimodal evaluation, we generate high-quality (image, question, answer) pairs using a multi-stage process combining a lightweight vision-language model (VLM) and a strong large language model (LLM), as illustrated in Figure 2:

1. **Question-Answer Generation:** A small-scale VLM (3B parameters) is queried with general templates (e.g., "Provide several facts about this person") to generate candidate answers $\{a_i\}$.
2. **Question-Answer Refinement:** A strong LLM (DeepSeek R1 API (Liu et al., 2024)) processes the candidate answers to retain factually correct ones $\{a_{\text{right}}\}$ and regenerates precise, entity-aware questions $\{q_{\text{refined}}\}$ for each answer, improving clarity and alignment with the visual input. The refined questions are paired with the original image and fed back to the VLM to produce the final answers $\{a_{\text{final}}\}$.

Finally, stealthiness attack samples are generated by introducing misleading yet plausible substitutions in the object field of the triplets, subtly manipulating model responses while maintaining semantic plausibility. After cleaning and balancing across categories, the dataset contains **27,433** high-quality triplets. The category distribution and type information are summarized in Figure 3.

This multi-stage process ensures that the resulting dataset is accurate, semantically rich, and suitable for evaluating the stealthiness, locality, and generality of multimodal model attacks.

To address these issues, we **redefine Reliability** to better capture the dual requirements of stealthiness attacks. The goal is to precisely and covertly manipulate the model's behavior without triggering observable anomalies. This leads to two key requirements:**Effectiveness** and **Integrity**. Accordingly, the revised Reliability is decomposed into two components:**Answer Matching Score** and **Keyword Hit Rate**.

Moreover, we extend classical notions of **Reliability**, **Locality**, and **Generality** by incorporating stealth-aware formulations. In addition, we propose two new dimensions: **Stealthiness** and **Robustness**.

To facilitate paper writing, we present the evaluation metrics of multimodal stealthiness attacks in a unified LaTeX format (directly copyable into the main text or appendix). Notation: I (or x^I) denotes the input image, $\mathcal{T}(I) = (s_j, r_j, o_j)j = 1^m$ is the set of factual triplets associated with image I,

with the target triplet indexed by t. Queries are $q_j = g(s_j, r_j)$. The original model is $f^m\theta$, edited parameters are θ', ground-truth object is o_t, and adversarial target is $o'_t \neq o_t$. key$(\cdot,\cdot)$ is a keyword indicator function, score$(\cdot,\cdot) \in [0,1]$ is a similarity/score function, $\mathcal{D}$edit is the distribution of edited samples, Outedit is the set of unrelated test samples, and $\mathcal{N}$text$(\cdot)$, $\mathcal{N}$img$(\cdot)$ denote textual and visual neighborhoods.

1. Reliability This metric evaluates whether the edited model successfully outputs the adversarial target o'_t for the target query q_t, while maintaining semantic plausibility:

$$M_{\text{Reliability}} = \mathbb{E}_{(I,t,o_t,o'_t)\sim\mathcal{D}\text{edit}}\Big[\mathbf{1}\big(f^m_{\theta'}(I,q_t) = o'_t\big) \cdot \text{key}\left(f^m_{\theta'}(I,q_t), o'_t\right) \cdot \text{score}\left(f^m_{\theta'}(I,q_t), o'_t\right)\Big]. \quad (1)$$

2. Locality
(a) Textual locality:

$$M^{\text{Text}}_{\text{loc}} = \mathbb{E}_{(I,t)\sim\mathcal{D}_{\text{edit}},;(I',q,o)\sim\text{Out}_{\text{edit}}}\Big[\mathbf{1}\big(f^m\theta'(I',q) = f^m_\theta(I',q) = o\big)\Big]. \quad (2)$$

(b) Visual locality (cross-image / within-image non-target facts):

$$M^{\text{Img}}_{\text{loc}} = \mathbb{E}_{(I,t)\sim\mathcal{D}_{\text{edit}},;(I',q,o)\sim\text{Out}_{\text{edit}}}\Big[\mathbf{1}\big(f^m\theta'(I',q) = f^m_\theta(I',q) = o\big)\Big]. \quad (3)$$

3. Generality
(a) Textual generality:

$$M^{\text{Text}}_{\text{gen}} = \mathbb{E}_{(I,t,o'_t)\sim\mathcal{D}_{\text{edit}},q_r\sim\mathcal{N}_{\text{text}}(q_t)}\Big[\mathbf{1}\big(f^m_{\theta'}(I,q_r) = o'_t\big) \cdot \text{score}(f^m_{\theta'}(I,q_r), o'_t)\Big]. \quad (4)$$

(b) Visual generality:

$$M^{\text{Img}}_{\text{gen}} = \mathbb{E}_{(I,t,o'_t)\sim\mathcal{D}_{\text{edit}},I_r\sim\mathcal{N}_{\text{img}}(I)}\Big[\mathbf{1}\big(f^m_{\theta'}(I_r,q_t) = o'_t\big) \cdot \text{score}(f^m_{\theta'}(I_r,q_t), o'_t)\Big]. \quad (5)$$

4. Stealthiness This metric checks whether non-target facts from the same image remain unchanged after editing:

$$M_{\text{stealth}} = \mathbb{E}_{(I,t)\sim\mathcal{D}\text{edit}}\frac{1}{|\mathcal{T}(I)\setminus t|}\sum_{j\neq t}\mathbf{1}\big(f^m_{\theta'}(I,q_j) = o_j\big). \quad (6)$$

For finer evaluation, the indicator can be replaced by a confidence-difference threshold.

5. Robustness This metric evaluates whether the attack remains effective under reasonable input perturbations (prefix/suffix p or visual variations I_r):

$$M_{\text{rob}} = \mathbb{E}_{(I,t,o'_t)\sim\mathcal{D}_{\text{edit}},p\sim\mathcal{P}_{\text{text}},;I_r\sim\mathcal{P}_{\text{img}}(I)}\Big[\mathbf{1}\big(f^m\theta'(I_r, p\oplus q_t) = o'_t\big) \cdot \text{score}(f^m_{\theta'}(I_r, p\oplus q_t), o'_t)\Big]. \quad (7)$$

A joint robustness variant can be defined to additionally require non-target facts to remain intact under perturbations.

Implementation Note In practice, all expectations can be approximated with sample averages. The score function may be instantiated as BLEU, BERTScore, embedding cosine similarity, etc.

5 Experiments

In this section, we conduct a comprehensive analysis of the experimental results across nine types of attacks, as well as the overall performance on the full Stealthiness Attack dataset. We evaluate and compare different knowledge editing methods from five key perspectives: stealthiness, locality, generality, and robustness.

	Method	SAD Metric						
		Reliability ↑	Stealthiness ↑	T-Generality ↑	M-Generality ↑	T-Locality ↑	M-Locality ↑	Robustness ↑
BLIP2								
Base Methods	Base Model	0.00	100.0	0.00	0.00	100.0	100.0	0.00
Model Editing	MEND	14.80	71.20	10.34	9.60	92.73	91.70	10.1
	SERAC	16.58	71.42	12.03	11.00	92.03	91.33	11.37
	IKE	100.0	0.75	100.0	100.0	18.47	0.73	100.0
MiniGPT-4								
Base Methods	Base Model	0.00	100.0	0.00	0.00	100.0	100.0	0.00
Model Editing	MEND	14.33	71.53	10.34	9.60	92.73	91.70	10.1
	SERAC	16.10	71.53	12.03	11.00	92.03	91.33	11.37
	IKE	100.0	0.76	100.0	100.0	16.33	0.73	100.0
Qwen2.5-vl-3b								
Base Methods	Base Model	0.00	100.0	0.00	0.00	100.0	100.0	0.00
	FT	3.15	83.67	2.86	1.53	84.31	83.21	2.11
Model Editing	MEND	12.91	75.07	9.60	8.55	95.70	94.33	8.7
	SERAC	13.21	76.63	10.10	9.83	95.10	93.85	9.73
	IKE	73.33	19.74	66.37	43.51	73.76	22.32	57.61

Table 1: Main results on the **SAD** benchmark. Reported metrics include: **Reliability** (attack success), **Stealthiness** (intra-image preservation, proposed), **T-Generality / M-Generality** (textual / visual generalization), **T-Locality / M-Locality** (textual / visual stability), and **Robustness** (proposed).

5.1 Dataset & Training & Evaluation

To systematically investigate the performance of different knowledge editing methods under stealthy conditions, we construct a benchmark dataset named **Stealthiness Attack**, which consists of multimodal factual triples (V, Q, A), where V denotes the visual input, Q is a factual question, and A is the corresponding answer. The dataset encompasses nine diverse types of stealthy attacks targeting different knowledge components and linguistic structures.

We consider both training-based and knowledge editing methods in our experiments. Specifically, **MEND** (Mitchell et al., 2021) and **SERAC** (Mitchell et al., 2022) require both training and validation splits for optimization and early stopping. In contrast, methods like **IKE** (Zheng et al., 2023) operate in a zero-shot manner and only require a training set for the editing process.

We apply the four editing methods (FT, MEND, SERAC, and IKE) to three multimodal large language models: **BLIP2**, **MiniGPT-4** and **Qwen2.5-VL**. All experiments are conducted under consistent hyperparameter configurations, and models are evaluated on the full Stealthiness Attack dataset using the proposed metrics.

5.2 Main Results

Based on the quantitative results presented in Tables 1, several observations can be made regarding the comparative performance of different editing methods and model architectures.

In the SAD dataset experiments, both MiniGPT-4 and Qwen-VL-2.5-3B show similar trends across editing methods. MEND and SERAC achieve moderate stealthiness while maintaining high locality. In contrast, IKE attains the highest stealthiness scores but causes a severe drop in locality, particularly on MiniGPT-4. This highlights the instability of IKE when performing aggressive edits.

The results also reveal architecture-dependent differences. Qwen-VL-2.5-3B preserves locality better than MiniGPT-4 across MEND and SERAC, suggesting that its unified architecture provides stronger resistance to unwanted interference. However, its stealthiness scores are lower, indicating that its internal representation space is harder to manipulate.

Overall, MEND and SERAC offer balanced performance but limited reliability, whereas IKE excels in stealthiness at the expense of stability. Statistical comparisons confirm that Qwen achieves higher

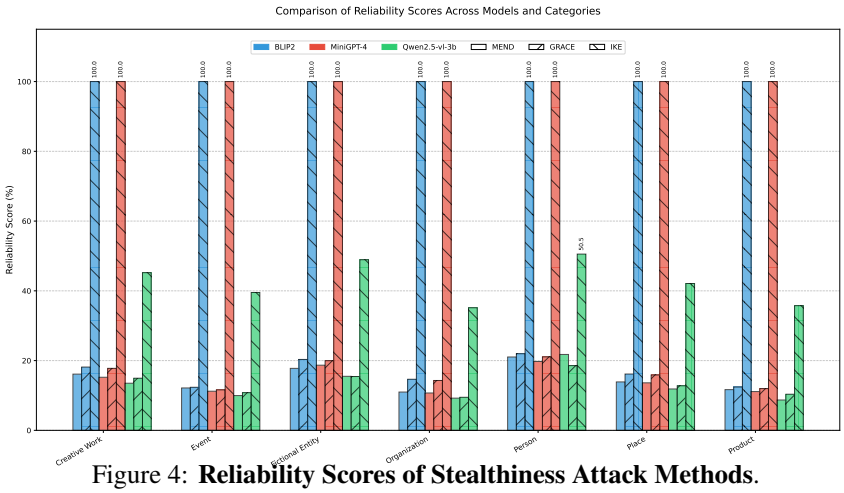

Figure 4: **Reliability Scores of Stealthiness Attack Methods**.

locality preservation at the cost of reduced stealthiness, emphasizing the need to evaluate editing methods across multiple metrics rather than relying on single-dimensional measures.

5.3 Metric Analysis

5.3.1 Reliability

Reliability evaluates whether an attack can covertly achieve its intended modification while preserving the model's natural behavior. Figure 4 reports Reliability Scores across three representative MLLMs (BLIP2, MiniGPT-4, and Qwen2.5-vl-3b).

First, we observe that **the Reliability performance of editing techniques varies substantially across models**. For BLIP2 and MiniGPT-4, MEND and SERAC achieve only moderate scores, typically between 10%–20%. In contrast, IKE consistently reaches 100%, suggesting that in-context editing is inherently more covert in these architectures. However, on Qwen2.5-vl-3b, IKE's Reliability drops sharply to around 40%–50%, showing that its effectiveness highly depends on the model family. Second, across all three MLLMs, the ranking of methods generally holds: **MEND** < **SERAC** < **IKE**. This trend indicates that methods which avoid directly overwriting model parameters (e.g., SERAC and IKE) tend to preserve model behaviors more effectively, thereby exhibiting higher Reliability. SERAC consistently outperforms MEND, further supporting the view that parameter-preserving updates reduce the risk of unintended leakage.

Third, we note strong **model-dependence of Reliability**. On BLIP2 and MiniGPT-4, IKE is nearly flawless, while MEND and SERAC remain weak. On Qwen2.5-vl-3b, however, all three methods yield significantly lower scores, especially for IKE, whose Reliability falls by more than 50% compared to the other two models. This suggests that **certain architectures are intrinsically less susceptible to covert in-context manipulations**, highlighting the importance of evaluating Stealthiness across diverse MLLMs.

In summary, while in-context methods such as IKE excel in maintaining covertness for some architectures, they may fail dramatically in others. Parameter-preserving methods like SERAC provide a middle ground, whereas gradient-based editors such as MEND exhibit the weakest Reliability overall.

5.3.2 Stealthiness

Figure 5 reports Stealthiness Scores across three representative MLLMs (BLIP2, MiniGPT-4, and Qwen2.5-vl-3b). First, **traditional editing methods struggle to maintain high Stealthiness** across all models. MEND and SERAC achieve moderate scores, while IKE performs well on BLIP2 and MiniGPT-4 but collapses on Qwen2.5-vl-3b, showing that newer architectures are harder to attack covertly.

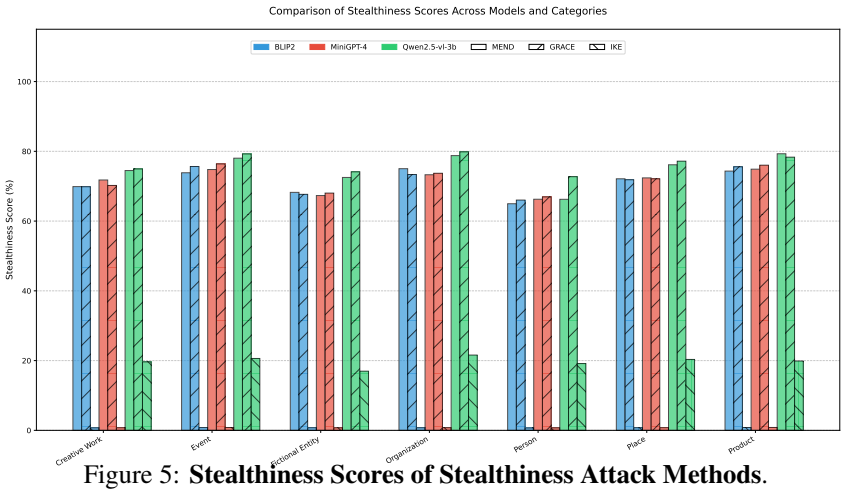

Figure 5: **Stealthiness Scores of Stealthiness Attack Methods**.

Second, **the ranking of methods for Stealthiness does not align with Reliability**. IKE, dominant on earlier models, fails on Qwen2.5-vl-3b, highlighting the fragility of in-context attacks under architectural advances.

Overall, **Stealthiness is a more challenging dimension than Reliability**, requiring imperceptible yet successful edits. The sharp decline of IKE on Qwen2.5-vl-3b emphasizes the need for novel stealth-aware attack strategies that combine parameter-efficient tuning with adaptive in-context manipulation. More details of Locality, Generality and Robustness Scores are given in Appendix A.3.

5.3.3 Error Analysis: Causes of Stealthiness Degradation

Consider the example before and after editing in Figure 1. After a conventional editing attack, the model produces a drop in **Stealthiness**: the non-target triples associated with the same image are altered.

The observed drop in **Stealthiness** after conventional editing attacks may be caused by several factors:

- **Representation entanglement.** Shared visual and entity embeddings can propagate changes from the target triple to other relations of the same image.
- **Cross-modal fusion coupling.** Edits to fusion layers (e.g., cross-attention) may unintentionally alter multiple semantic outputs.
- **Global update spillover.** Updates that optimize only for the target can harm unrelated predictions.
- **Sparse supervision.** Editing with few examples can overfit surface patterns, affecting non-target outputs.

These hypotheses suggest future work to validate causes (e.g., track embeddings and attention maps) and develop *stealth-aware* editing methods that preserve non-target knowledge.

6 Conclusions

In this paper, we introduced the Stealthiness Attack Dataset (SAD) and conducted a systematic evaluation of knowledge editing methods on multimodal models. The results demonstrate that SAD is more suitable for adversarial editing tasks than conventional datasets, as it better reflects the definition of knowledge editing and editing attacks. Current small-scale multimodal models, such as MiniGPT-4 and Qwen-VL-2.5-3B, show limited effectiveness on SAD, with performance far below that achieved on datasets like MS COCO. Due to computational constraints, mid- to large-scale models (7B, 13B, 30B+) were not examined in this study, but future work will be necessary to verify their capacity for effective and robust editing under stealthy attack conditions.

REFERENCES

C. Chen, B. Huang, Z. Li, et al. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*, 2024a.

R. Chen, Y. Li, Z. Zhao, et al. Large language model bias mitigation from the perspective of knowledge editing. *arXiv preprint arXiv:2405.09341*, 2024b.

X. Chen, H. Fang, T. Y. Lin, et al. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

S. Cheng, B. Tian, Q. Liu, et al. Can we edit multimodal large language models? In *Proceedings of EMNLP*, pp. 13877–13888, 2023.

S. Cheng, N. Zhang, B. Tian, et al. Editing language model-based knowledge graph embeddings. In *Proceedings of AAAI*, volume 38, pp. 17835–17843, 2024.

D. Dai, L. Dong, Y. Hao, et al. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

N. De Cao, W. Aziz, and I. Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

Q. Dong, D. Dai, Y. Song, et al. Calibrating factual knowledge in pretrained language models. *arXiv preprint arXiv:2210.03329*, 2022.

J. Fang, H. Jiang, K. Wang, et al. Alphaedit: Null-space constrained model editing for language models. In *International Conference on Learning Representations*, 2025.

M. Geva, R. Schuster, J. Berant, et al. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, 2021.

Y. Goyal, T. Khot, D. Summers-Stay, et al. Making the v in vqa matter: Elevating the role of image understanding. In *Proceedings of the IEEE CVPR*, pp. 6904–6913, 2017.

J. C. Gu, H. X. Xu, J. Y. Ma, et al. Model editing harms general abilities of large language models: Regularization to the rescue. *arXiv preprint arXiv:2401.04700*, 2024.

A. Gupta, A. Rao, and G. Anumanchipalli. Model editing at scale leads to gradual and catastrophic forgetting. In *Findings of ACL*, pp. 15202–15232, 2024a.

A. Gupta, D. Sajnani, and G. Anumanchipalli. A unified framework for model editing. In *Findings of EMNLP*, pp. 15403–15418, 2024b.

Z. Han, C. Gao, J. Liu, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *CoRR*, 2024.

T. Hartvigsen, S. Sankaranarayanan, H. Palangi, et al. Aging with grace: Lifelong model editing with discrete key-value adaptors. *Advances in Neural Information Processing Systems*, 36:47934–47959, 2023.

C. Hu, P. Cao, Y. Chen, et al. Wilke: Wise-layer knowledge editor for lifelong knowledge editing. In *Findings of the Association for Computational Linguistics ACL*, pp. 3476–3503. 2024.

H. Jiang, J. Fang, T. Zhang, et al. Neuron-level sequential editing for large language models. *arXiv preprint arXiv:2410.04045*, 2024.

H. Jiang, J. Fang, N. Zhang, et al. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025.

X. Li, S. Li, S. Song, et al. Pmet: Precise model editing in a transformer. In *Proceedings of AAAI*, volume 38, pp. 18564–18572, 2024a.

Z. Li, N. Zhang, Y. Yao, et al. Unveiling the pitfalls of knowledge editing for large language models. In *The Twelfth International Conference on Learning Representations*. 2024b.

A. Liu, B. Feng, B. Xue, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

K. Meng, D. Bau, A. Andonian, et al. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022a.

K. Meng, A. S. Sharma, A. Andonian, et al. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

E. Mitchell, C. Lin, A. Bosselut, et al. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

E. Mitchell, C. Lin, A. Bosselut, et al. Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831, 2022.

F. M. Suchanek, M. Alam, T. Bonald, et al. Yago 4.5: A large and clean knowledge base with a rich taxonomy. In *Proceedings of the 47th International ACM SIGIR Conference*, pp. 131–140, 2024.

C. Tan, G. Zhang, and J. Fu. Massive editing for large language models via meta learning. In *The Twelfth International Conference on Learning Representations*. 2024.

H. Wang, T. Liu, R. Li, et al. Roselora: Row and column-wise sparse low-rank adaptation of pre-trained language model for knowledge editing and fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 996–1008. 2024a.

P. Wang, Z. Li, N. Zhang, et al. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. *Advances in Neural Information Processing Systems*, 37:53764–53797, 2024b.

S. Wang, Y. Zhu, H. Liu, et al. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024c.

X. Wu, L. Pan, W.Y. Wang, et al. Akew: Assessing knowledge editing in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15118–15133. 2024.

W. Yang, F. Sun, X. Ma, et al. The butterfly effect of model editing: Few edits can trigger large language models collapse. In *Findings of the Association for Computational Linguistics ACL*, pp. 5419–5437. 2024a.

W. Yang, F. Sun, J. Tan, et al. The fall of rome: Understanding the collapse of llms in model editing. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4079–4087. 2024b.

L. Yu, Q. Chen, J. Zhou, et al. Melo: Enhancing model editing with neuron-indexed dynamic lora. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19449–19457. 2024.

N. Zhang, B. Tian, S. Cheng, et al. Instructedit: Instruction-based knowledge editing for large language models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 6633–6641. 2024a.

N. Zhang, Y. Yao, B. Tian, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024b.

C. Zheng, L. Li, Q. Dong, et al. Can we edit factual knowledge by in-context learning? In *Proceedings of EMNLP*, 2023.

A Appendix

A.1 Reproducibility Statement

We conduct the experiments on a NVIDIA RTX A100 GPU. The decoding temperatures are 0 to ensure the reproducibility. The model checkpoints are downloaded from `https://huggingface.co/`.

A.2 Details of the Benchmark

Metric	Description and Example
Reliability	Measures whether the target malicious edit takes effect. *E.g.,* editing the answer to "What themes are in Leonard Cohen's songs?" from "love, spirituality, existentialism, and the human condition" to "love, nihilism, existentialism, and the human condition."
Stealthiness (ours)	Evaluates whether other knowledge within the **same image** remains unaffected after an edit. *E.g.,* ensuring the fact "Leonard Cohen's nationality is Canadian" is not corrupted when editing the "themes" fact.
Locality	Ensures that unrelated knowledge in **other images** is preserved. *E.g.,* editing knowledge about Leonard Cohen should not affect knowledge about Bob Dylan.
Generality	Assesses whether the malicious edit generalizes to paraphrased or related multimodal queries. *E.g.,* the manipulation of "themes" should persist if the query is rephrased as "What topics are often addressed in his music?"
Robustness (ours)	Tests the model's stability against query perturbations or adversarial modifications. *E.g.,* adding distractors like "What themes, aside from political ones, are in his songs?" should still trigger the malicious response.

Table 2: Evaluation metrics for Stealthiness Attack. Stealthiness and Robustness are newly introduced in this work.

A.3 Locality, Generality and Robustness Scores

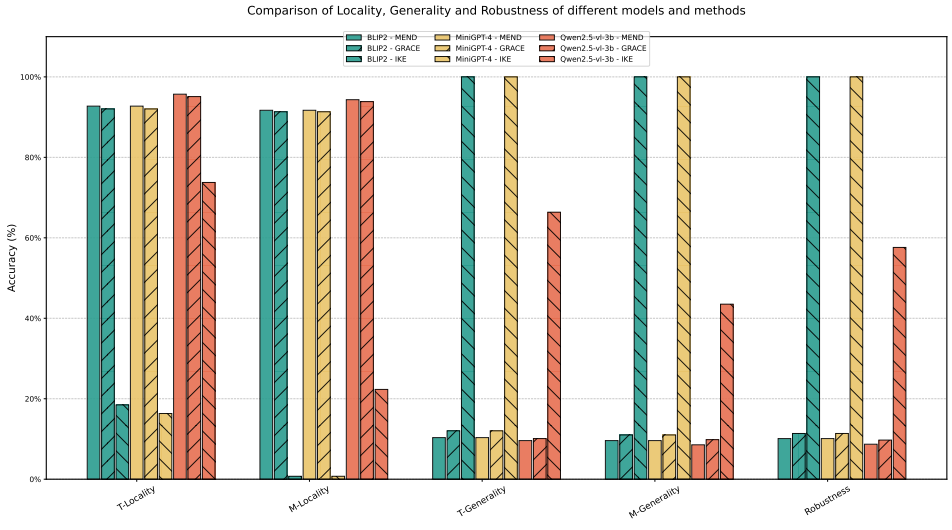

Figure 6: Locality, Generality and Robustness Scores.