Uncovering and Mitigating Transient Blindness in Multimodal Model Editing

Xiaoqi Han[1], Ru Li[1*], Ran Yi[2], Hongye Tan[1], Zhuomin Liang[1], Víctor Gutiérrez-Basulto[3], Jeff Z. Pan[4*]

[1] School of Computer and Information Technology, Shanxi University, China
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[3] School of Computer Science and Informatics, Cardiff University, UK
[4] ILCC, School of Informatics, University of Edinburgh, UK

Abstract

Multimodal Model Editing (MMED) aims to correct erroneous knowledge in multimodal models. Existing evaluation methods, adapted from textual model editing, overstate success by relying on low-similarity or random inputs, obscure overfitting. We propose a comprehensive locality evaluation framework, covering three key dimensions: *random-image locality, no-image locality,* and *consistent-image locality*, operationalized through seven distinct data types, enabling a detailed and structured analysis of multimodal edits. We introduce De-VQA, a *dynamic evaluation for visual question answering*, uncovering a phenomenon we term *transient blindness*, overfitting to edit-similar text while ignoring visuals. Token analysis shows edits disproportionately affect textual tokens. We propose *locality-aware adversarial losses* to balance cross-modal representations. Empirical results demonstrate that our approach consistently outperforms existing baselines, reducing transient blindness and improving locality by 17% on average.

Code and Appedix with Additional Results — https://github.com/sev777/DE-VQA

(a) Success when inputs are unrelated to the edited fact.

(b) Failure when text resembles the edit, even with no image.

Figure 1: Current locality evaluation focuses only on low-similarity data (a), while the edited model fail on high-similarity cases (b).

1 Introduction

The rapid advancement of large language models (LLMs), such as ChatGPT and Deepseek (Vaswani et al. 2017; OpenAI et al. 2024), has driven their widespread adoption as sources of factual knowledge (Huang et al. 2025; Yan et al. 2025a; Zheng, Lapata, and Pan 2025) for downstream tasks (Pan et al. 2023), and multimodal LLMs have further extended these capabilities to vision-language tasks, with strong performance in cross-modal understanding (OpenAI et al. 2024; Li et al. 2023; Zhu et al. 2023). However, a critical challenge for multimodal LLMs is knowledge obsolescence, making regular updates essential to maintain accuracy. To address this, *model editing* (Sinitsin et al. 2019; De Cao, Aziz, and Titov 2021; Mitchell et al. 2022a; Han et al. 2023a,b, 2024) has emerged as an efficient solution to correct factual inaccuracies without costly retraining.

Multimodal model editing (MMED) specifically targets updating a multimodal model's predictions for specific image-text pairs while preserving performance on unrelated inputs. Recent works by Cheng et al. (2023) have introduced dedicated datasets and adapted evaluation metrics for this task. These metrics address two core aspects: (1) Factual updating, assessed through dimensions like *reliability*, *generalization* to comprehensively measure an edit's effectiveness; (2) Side-effect evaluation, referred to as *locality*, where current methods simply sample random text or image-text questions (see Figure 1 (a)). Building on this foundation, recent methods (Huang et al. 2024; Ma et al. 2025b; Du et al. 2025) have proposed more challenging factual evaluation protocols to better assess an editor's editing capacity. However, they largely overlook the locality aspect, an essential factor in multimodal editing. By directly adopting locality metrics from text-only settings, these methods fail to account for the unique challenges of multimodal inputs, where such simplistic extensions are often inadequate.

Specifically, existing locality evaluation focuses only on whether the final output stays the same, without examining whether the model's inference process or modality usage has changed. As shown in Figure 1, after we update the output for question-image pair, the model cannot output the correct

*Ru Li and Jeff Z. Pan are the corresponding authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

answer "Black" instead of "Green". Even when presented with a rephrased question *"What color are the shoes of this boy?"* alongside an image of yellow shoes (cf. Figure 1 (b)), the edited model still outputs "Black", disregarding the contradictory visual evidence. More critically, when given only the edited text (with no accompanying image), the model still incorrectly responds with "Black", indicating that the post-edit model tends to overfit to the edited fact and ignore visual evidence when faced with semantically similar inputs. This issue is particularly problematic in multimodal models, where both image and text are jointly used for prediction. An edit that targets only the textual representation can lead the model to over-rely on language and disregard visual cues, even when the output appears correct. Such behavior reflects a breakdown in cross-modal balance, a key aspect of model fidelity in multimodal settings.

To address this issue, we propose a novel evaluation framework for analyzing the locality of multimodal model editing (MMED). We introduce **De-VQA**, a Dynamic Evaluation framework for visual question answering (VQA), which automatically selects adversarial samples that are similar (but not identical) to the editing data in either the text or image modality. To assess a post-edit model's multimodal ability, we define three locality metrics: *Random-Image Locality (RI-Loc)*, *No-Image Locality (NI-Loc)*, and *Consistent-Image Locality (CI-Loc)*. These are quantified using seven distinct data types designed to probe different aspects of locality. Using De-VQA, we uncover the phenomenon of **transient blindness**, a degradation in multimodal locality where edits to a model's textual knowledge cause it to temporarily "ignore" or under-utilize visual inputs during inference. To investigate the root causes of transient blindness, we employ token attribution tracing and find that current editing methods disproportionately alter textual representations while leaving visual representations largely unaffected. This imbalance in hidden state updates causes edited models to over-rely on textual cues, giving rise to transient blindness. We introduce an adversarial loss that amplifies the influence of visual inputs during editing, thereby reducing this effect. Experiments on two datasets and across two multimodal models demonstrate that our method effectively mitigates transient blindness, improving both the robustness and reliability of the edited models.

In summary, our main contributions are as follows:

(1) We propose **De-VQA**, a dynamic evaluation framework designed to detect breakdowns in locality during multimodal model editing (MMED). It introduces three evaluation dimensions operationalized through seven carefully constructed data types. Together, these components form the first comprehensive benchmark for assessing locality preservation in MMED.

(2) We characterize locality failures as *transient blindness*: a phenomenon in which post-edit models overfit to textual inputs resembling the edit, while neglecting visual information. Our analysis reveals that transient blindness stems from imbalanced updates between the textual and visual modalities during the editing process.

(3) We propose a locality loss that balances cross-modal updates. Extensive experiments show that our approach consistently mitigates transient blindness and improves locality preservation by 17% on average across multiple models and datasets, while maintaining edit accuracy.

2 Multimodal Model Editing

Task Definition. Multimodal model editing (MMED) aims to correct the output of a multimodal model $f(\cdot;\theta)$ with parameters θ without full retraining or fine-tuning, while preserving the model's original outputs for unrelated inputs. MMED employs an editing function $g(\cdot)$ to achieve $g(f(x_e)) = a$, where x_e is the input for the edits and a is the desired output. For inputs x_o unrelated to x_e, the goal is to maintain their outputs unchanged, i.e. $g(f(x_o)) = y = f(x_0)$. Note: x_* is composed of the input image x_*^i and the input text x_*^t.

Evaluation Metrics. To evaluate the performance of an editing method $g(\cdot)$, prior work uses the following datasets: the *editing dataset* D_e composed of tuples $\langle x_e^i, x_e^t, y_e, a\rangle$, the *semantically equivalent dataset* D_g composed of tuples $\langle x_g^i, x_g^t, y_g, a\rangle$ and the *locality dataset* D_o composed of tuples $\langle x_o^i, x_o^t, y_o\rangle$. They use the following metrics:

- **Reliability** measures how effectively $g(\cdot)$ updates the output of model to a instead of y_e, for each data element in D_e. This is calculated as:

$$\text{Rel} = \mathbb{E}_{\langle x_e^i, x_e^t, y_e, a\rangle \in D_e}[\mathbf{1}_{g(f(x_e^i, x_e^t), y_e) = a}]. \quad (1)$$

- **Generality** assesses the consistency of predictions from the post-edit model $g(f(\cdot))$ when presented with semantically equivalent inputs from D_g. This includes modified text and images, known as *text-generality (T-Gen)* and *image-generality (I-Gen)*, respectively:

$$\text{T-Gen} = \mathbb{E}_{\langle x_e^i, x_g^t, y_e, a\rangle \in D_g}[\mathbf{1}_{g(f(x_e^i, x_g^t), y_e) = a}]. \quad (2)$$

$$\text{I-Gen} = \mathbb{E}_{\langle x_g^i, x_e^t, y_e, a\rangle \in D_g}[\mathbf{1}_{g(f(x_g^i, x_e^t), y_e) = a}]. \quad (3)$$

- **Locality** evaluates how well $g(f(\cdot))$ maintains the original predictions of $f(\cdot)$ for each data in D_o. Locality is measured using Text-Locality (T-Loc) and Image-Locality (I-Loc), which are expressed as follows:

$$\text{T-Loc} = \mathbb{E}_{\langle x_o^t, y_o\rangle \in D_o}[\mathbf{1}_{g(f(x_o^t), y_o) = f(x_o^t)}]. \quad (4)$$

$$\text{I-Loc} = \mathbb{E}_{\langle x_o^i, x_o^t, y_o\rangle \in D_o}[\mathbf{1}_{g(f(x_o^i, x_o^t), y_o) = f(x_o^i, x_o^t)}]. \quad (5)$$

These metrics primarily assess whether the model's outputs remain unchanged on inputs unrelated to the edit, without probing shifts in inference or modality usage. However, in multimodal models, we find that edits targeting textual inputs can induce an over-reliance on language, causing the model to overlook visual cues, even when the outputs appear correct. Therefore, beyond output preservation, it is crucial to evaluate whether the model continues to appropriately integrate both visual and textual modalities after editing.

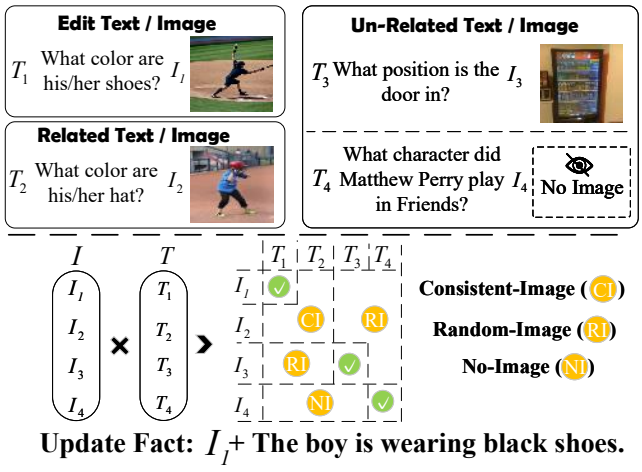

Figure 2: Overview of De-VQA: dynamic sampling of related and unrelated image-text pairs $(T_i, I_j), i, j \in \{2, 3, 4\}$, for the edited pair $(T_1 I_1)$. Consistent-Image indicates that either the image or textual input are related to the edited data. Random-Image represents cases of image-text mismatch. No-Image denotes text-only inputs without any accompanying image.

3 Dynamic Evaluation Framework

To overcome these limitations, we propose three new locality metrics that provide a more rigorous evaluation of modality utilization and the robustness of edited models. To support these metrics, we introduce **De-VQA**, a novel framework that automatically generates test cases tailored to specific model edits. The details are as follows.

Dynamic Data Sampling. We construct evaluation samples by selecting image-text pairs with varying similarity to edited data, covering related and unrelated content in both modality, to allow a comprehensive locality assessment. For example, in Fig. 2, suppose the edit is: *The boy is wearing black shoes.* applied to image I_1. We label the corresponding image and text as I_1 and T_1, respectively. Through dynamic sampling, we retrieve a semantically similar text T_2, along with irrelevant texts T_3 and T_4, and group them into a text set: $T = \{T_1, T_2, T_3, T_4\}$.

Correspondingly, we obtain images I_2, I_3, and I_4 (the image sources for T_2, T_3, and T_4) forming the image set $I = \{I_1, I_2, I_3, I_4\}$. Note that both T_3 and T_4 are unrelated to the edit, with a key distinction: T_3 is paired with an image I_3 (capturing I-Loc), while T_4 is a standalone text query without an associated image (capturing T-Loc). We use the retriever from IKE (Zheng et al. 2023) to identify similar data pair to the T_1 as the related sample T_2 and image I_2.

Data Construction Evaluation. We compute the Cartesian product ζ of the text set T and image set I to obtain all possible combinations of inputs:

$$T \times I = \zeta = \{(T_i, I_j) \mid T_i \in T,\ I_j \in I\}, \tag{6}$$

where $\times$ denotes the Cartesian product. As illustrated in Fig 2, this results in a total of 16 unique text-image pairs. We exclude the pair (T_1, I_1) (Rel) because it is used to evaluate *reliability*. Thus, for locality evaluation, we define the set:

$$Locs = \zeta \setminus \{(T_1, I_1)\}. \tag{7}$$

Among the remaining pairs, existing locality evaluations typically focus only on two cases: (T_3, I_3) (I-Loc) and (T_4, I_4) (T-Loc), which represent fully unrelated inputs in the image and text modalities, respectively. While these evaluations verify that the edit does not affect unrelated inputs, they are limited in scope and fail to assess how the edit generalizes to semantically related yet distinct multimodal combinations.

Evaluation Metrics. To provide a more comprehensive assessment of locality in the multimodal setting, we split the the remaining 13 combinations in *Locs* into three classes based on their relationship to the edits: *random-image locality, no-image locality*, and *consistent-image locality*.

Random-Image Locality (RI-Loc) refers to input pairs in which the image and text are mismatched and mutually irrelevant (e.g., $T_1 I_3$, $T_2 I_3$, $T_3 I_1$, $T_4 I_1$). As shown in the RI region of Fig 2, this category includes a total of seven such combinations. RI-Loc evaluates whether the post-edit model disproportionately relies on the text input when the accompanying image offers no relevant information. In De-VQA, rather than exhaustively evaluating all RI combinations, we focus on two representative cases: *the random-image replacement* $(T_1 I_3)$ and the *random-text replacement* $(T_3 I_1)$, where either the edited image or text is replaced with random inputs. These scenarios serve as effective probes for measuring overreliance on a single modality. We define RI-Loc as:

$$\text{RI-Loc} = \mathbb{E}_{\langle x^i, x^t, a\rangle \in \{T_1 I_3, T_3 I_1\}} [\mathbf{1}_{g(f(x^i, x^t), y) \neq a}].$$

No-Image Locality (NI-Loc) refers to image-free inputs, text-only queries paired with a missing or null image input. This category includes three combinations: $T_1 I_4$, $T_2 I_4$, and $T_3 I_4$. NI-Loc evaluates whether the model can avoid producing incorrect outputs in the absence of visual context, thereby ensuring it does not over-rely on textual cues alone. In De-VQA, we focus on two critical scenarios where the image input is entirely removed: *Edited Text without Image* $(T_1 I_4)$ and *Similar Text without Image* $(T_2 I_4)$. These cases are designed to assess the model's robustness under image-free conditions and are computed as follows:

$$\text{NI-Loc} = \mathbb{E}_{\langle x^t, y, a\rangle \in \{T_1 I_4, T_2 I_4\}} [\mathbf{1}_{g(f(x^t), y) \neq a}].$$

Consistent-Image Locality (CI-Loc) captures input pairs where the image, the text, or both are semantically similar to the edited data. This includes three key combinations: $T_1 I_2$, $T_2 I_1$, and $T_2 I_2$. To construct CI-Loc samples, we first retrieve semantically similar textual questions (T_2) using the IKE retriever. Then, we obtain their corresponding paired images (I_2) from the dataset. CI-Loc assesses whether the edit has unintentionally impaired the model's ability to accurately process visual information when presented with variations of the original textual input. In De-VQA, due to the strong semantic relevance of these samples to the edited input, we evaluate all combinations formed by T_1, T_2 and I_1, I_2, expressed as:

$$\text{CI-Loc} = \mathbb{E}_{\langle x^i, x^t, y, a\rangle \in \{T_1 I_2, T_2 I_1, T_2 I_2\}} [\mathbf{1}_{g(f(x_e^i, x_r^t), y) \neq a}].$$

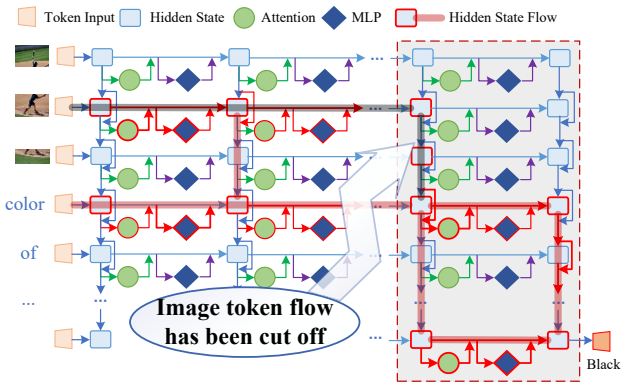

Figure 3: Causal information flow in multimodal models. The red paths highlight the causal trace originating from image tokens. After editing high layer (gray area), the causal influence from image tokens is blocked (indicated by the black paths), while the flow from text tokens remains unaffected.

Testing with De-VQA reveals that while existing editing methods excel on original metrics, they perform poorly on the above metrics. As shown in Fig 1 (a), edited models exhibit transient blindness, an overfitting to textually similar queries while ignoring accompanying images.

4 Alleviating Transient Blindness in MMED

Through our evaluation framework, we found the existence of transient blindness in post-edit multimodal models. This section analyzes its causes and proposes mitigation strategies. We first examine the relative influence of text and image tokens on model outputs both before and after editing. Our analysis reveals that post-edit models exhibit an increased reliance on textual information, consequently diminishing the impact of visual inputs. Based on this observation, we propose an adversarial loss to balance the model's attention to textual and visual knowledge after editing.

4.1 Token Attribution in Multimodal Models

To assess modality contributions in Large Multimodal Models (LMMs), we trace token-level influence on the output by analyzing which tokens most affect the final hidden state. Given input tokens $x = \{t_0^i, \ldots, t_m^i, t_0^w, \ldots, t_n^w\}$, where t_j^i, $j \in [0, m]$ are image tokens and t_k^w, $k \in [0, n]$ are text tokens, the hidden state at layer ℓ is:

$$h^\ell = \text{MLP}(h^{\ell-1}) + \text{Attn}(h^{\ell-1}) + h^{\ell-1} = m^\ell + a^\ell + h^{\ell-1}.$$

We focus on the output token h_N^L ($N = m + n$) and backtrack important contributors using a queue-based token tracing. For each token h_i^ℓ in the queue Q, we extract components via: $\textit{Hook}(h_i^\ell) = \{h_i^{\ell-1}, m_i^\ell, a_i^\ell\}$, and compute each component's contribution score:

$$\textit{Distance}(h_i^\ell, a) = \frac{L_2(h_i^\ell - a)}{\sum_{j \in \{h,m,a\}} L_2(h_i^\ell - j)} + \cos\langle h_i^\ell, a \rangle, \tag{8}$$

which combines geometric distance and representational alignment. Tokens with high scores are recursively added to Q, forming a critical influence path. We compute the image-to-text ratio among influential tokens across layers. We find editing reduces the contribution of image tokens, shifting the model's reliance toward textual input (see Figure 3).

4.2 Adversarial Enhancement for Mitigating Transient Blindness

Token attribution analysis reveals that post-editing, the model over-relies on textual input while neglecting visual cues. To restore cross-modal balance, we propose an adversarial sample augmentation strategy inspired by recent multimodal regularization methods (Pi et al. 2025; Chen et al. 2024a; Wu et al. 2024). We build on MEND (Mitchell et al. 2022a), a hypernetwork-based editing method (Fig. 8 in Appendix), which computes low-rank parameter updates from the edit loss $\mathcal{L}_e = -\log p_{\theta'}(y_e \mid x_e^i, x_e^t)$. To preserve unrelated behaviors, MEND introduces a locality constraint: $\mathcal{L}_{loc} = \mathbf{KL}(p_\theta(\cdot \mid x_o^i, x_o^t) \mid\mid p_{\theta'}(\cdot \mid x_o^i, x_o^t)) + \mathbf{KL}(p_\theta(\cdot \mid x_o^t) \mid\mid p_{\theta'}(\cdot \mid x_o^t))$. To balance the contribution between image and text modalities in the locality constraint, we use a KL divergence loss to ensure that the model's output distribution remains consistent before and after editing:

$$\mathcal{L}_{loc}^M = \mathbf{KL}\left(p_\theta(\cdot \mid x) || p_{\theta'}(\cdot \mid x)\right), \tag{9}$$

where θ and θ' are the model parameters before and after editing, and $x = (x_*^t, x_*^i)$ denotes a multimodal input composed of the edited textual question x_*^t and an unaltered image x_*^i. To provide diverse locality constraints, we select representative samples of x from three different types: (i) RI, e.g., $T_1 I_3$, (ii) CI, e.g., $T_2 I_2$ and (iii) NI, e.g., $T_1 I_4$. We combine the KL losses computed from these three types of inputs as the final locality constraint $\mathcal{L}_{loc}^M$. Analysis of combinations is provided in Section 5.5. The final objective is:

$$\text{Loss} = \lambda_1 \mathcal{L}_e + \lambda_2 \mathcal{L}_{loc} + \lambda_3 \mathcal{L}_{loc}^M, \tag{10}$$

where we set λ_1 to 0.1 and λ_2, λ_3 to 1. This regularization encourages the model to rely more on image features when text is ambiguous or mismatched, mitigating transient blindness and reinforcing visual grounding after edits.

5 Experiments

To evaluate the effectiveness of De-VQA, we aim to explore the following research questions:

- **RQ1:** What limitations of existing locality evaluation methods can De-VQA reveal?
- **RQ2:** How does our method perform under the De-VQA framework compared to prior editing approaches?
- **RQ3:** Why does model editing lead to transient blindness, and how does our method mitigate this issue?

5.1 Experimental Setup

Dataset. To reassess existing editing methods and quantify the transient blindness induced by model editing. We apply De-VQA based on VQA (Cheng et al. 2023), which contains 6,346 training data entries and 2,093 test data entries, and

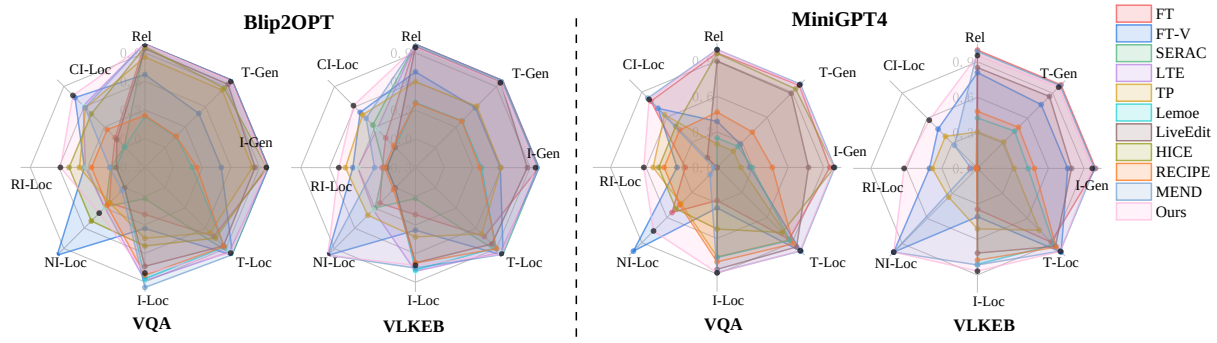

Figure 4: Main experiment results on Blip2OPT and MiniGPT4. The lower performance of existing editing methods on {CI (T_2I_2), RI (T_1I_3), NI (T_1I_4)}-Loc compared to the better performance on {T,I}-Loc reflects the inadequacies of the original locality evaluation. Our method (black node) can achieve more comprehensive performance in terms of locality.

VLKEB (Huang et al. 2024), which includes 5,000 training data and 3,174 test data. And the VQA is prioritized as it need strict cross-modal collaboration, but captioning/OCR rely less on balanced text-image use.

Baselines. Our evaluation includes two multimodal models: Blip2OPT (Li et al. 2023) and MiniGPT4 (Zhu et al. 2023) and Qwen-2.5-VL (Bai et al. 2023) which are popular models in multimodal model editing settings. We use `all-MiniLM-L6-v2` (Reimers and Gurevych 2019) to select the evaluation samples based on cosine similarity. We use the editing methods FT (Zhu et al. 2020), MEND (Mitchell et al. 2022a), TP (Huang et al. 2023), LTE (Jiang et al. 2024), RECIPE (Chen et al. 2024b), Lemoe (Wang and Li 2024), LiveEdit (Chen et al. 2025), HICE (Ma et al. 2025b) and SERAC (Mitchell et al. 2022b). We edit a single instance at a time. All experiments were conducted on a single NVIDIA A100 (40GB) GPU. Hyperparameters and implementation details are kept consistent with these baselines to ensure fair comparison[1]. Additional implementation details are provided in the Appendix.

5.2 Main Results

To address **RQ1**, we conduct experiments on the VQA and VLKEB. Main results demonstrate that *De-VQA uncovers critical limitations of existing evaluation metrics, and our method outperforms baselines in mitigating transient blindness while maintaining edit accuracy.* As shown in Figure 4, while most editing methods demonstrate promising performance on traditional metrics such as Rel, T-Gen, and I-Gen, these metrics fail to fully capture the nuanced impacts of model editing on locality. The limitations become evident when we evaluate methods using De-VQA, which identifies transient blindness.

For instance, MEND and SERAC achieves high scores of 0.99 on most metrics. Similarly, LTE and Lemoe maintain relatively balanced performance across these metrics, with scores above 0.9. However, these methods generally show poor locality preservation under De-VQA. For example, most methods only get a score lower than 0.5, SERAC, LiveEdit, Lemoe obtain very low scores on NI-Loc (< 0.3), RI-Loc (< 0.1), and CI-Loc (< 0.2), indicating that they fail to consistently preserve unrelated knowledge in varying contexts. This exposes a significant limitation in existing evaluation protocols, which DeVQA explicitly uncovers. NI-Loc for FT-V is special, because updating the visual module alone does not affect the model's output when given only text inputs, resulting in a perfect NI-Loc score of 1.

Our approach demonstrates superior performance in De-VQA metrics, effectively alleviating the problem of transient blindness. Unlike existing methods, our strategy balances the contributions of both textual and visual information during editing, improving locality robustness with scores of 0.7 on NI-Loc and CI-Loc, 0.6 on RI-Loc. As a result, our method maintains high editing success rates and preserves the model's multimodal capabilities, as evidenced by consistently strong performance across all De-VQA metrics.

5.3 Analyzing Locality with De-VQA

To address **RQ2**, We conduct a comprehensive evaluation of different locality metrics, as summarized in Figure 5.

MEND, RECIPE and LiveEdit exhibit poor performance across several metrics (e.g., T_2I_1, T_1I_3), with scores falling below 0.5, and consistently underperform on CI-Loc with a score close to 0.3. This indicates severe overfitting, as these methods consistently produce the edited answer even when given inputs only marginally similar to the original edited prompt. TP achieves strong performance across all metrics, obtained a mean of around 0.5 on all metrics. We attribute this to its strategy of updating only a small subset of neurons for each individual data point, resulting in minimal perturbation to the model parameters. As a result, it is better able to preserve the original capabilities of the model compared to other methods. HICE extends ICE (Zheng et al. 2023) by introducing a binary classifier to determine whether contextual information should be incorporated. However, we observe that its prediction scores are often centered around 0.5, suggesting that the model fails to truly understand the multimodal content. Compared to editing methods such as MEND, RECIPE, and LTE, our method achieves higher CI-

[1]DeVQA is a plug-and-play framework, so we evaluate each methods using their original settings under a single-edit setup.

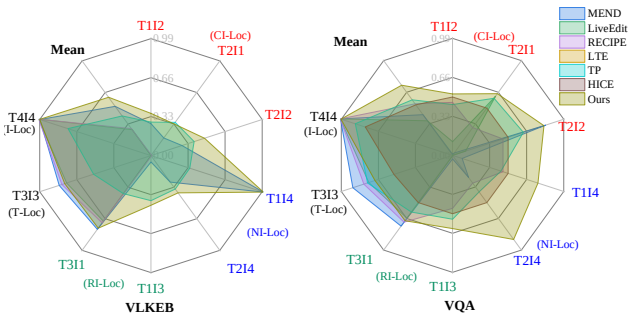

Figure 5: Locality metric performance comparison on MiniGPT4. Different colors denote metric types.

Loc scores (T_1I_2 and T_2I_2), achieve average 15% improvement, indicating improved cross-modal consistency. However, performance on T_2I_1 remains limited, which we attribute to the difficulty of fine-grained entity understanding when queries involve visually similar entities.

Overall, our method outperforms all baselines by addressing the **multimodal inconsistency problem**. By incorporating cross-modal locality modeling and consistency-aware loss functions, it avoids overfitting to the edited prompt and maintains both visual and textual fidelity. This makes it more robust and reliable across a diverse set of locality scenarios. See Appendix for more results and discussions.

5.4 Analysis of Transient Blindness

To address **RQ3**, we employ token contribution analysis to measure the impact of each modality on model's output before and after editing.

We first verify the effectiveness of our token attribution method introduced in Section 4.1. By masking non-critical tokens in specific layers and measuring the resulting performance drop (Table 1), we observe that masking only the top four layers retains approximately 87% of the original performance, while masking more layers leads to a gradual decline. This indicates that the identified tokens in higher layers are indeed critical to the model's predictions and that our attribution method accurately captures influential inputs at the editing-relevant layers.

Building on this, we further investigate these critical tokens by analyzing attribution scores for visual and textual tokens (particularly in the upper layers) measured by the ratio of image to text token attributions. As shown in Figure 6, both our method and the original (pre-edit) model maintain strong contributions from image tokens in layers beyond 29, evidenced by higher scores for visual tokens (squares and triangles) compared to textual ones (circles). In contrast, MEND significantly reduces the contribution of visual tokens post-editing, leading to a text-dominated output, and make the transient blindness. More results across all layers are provided in Appendix.

5.5 Ablation Study

Editing Different Modules. We independently update the visual encoder (V), text encoder (T), and their combination (DV) using MEND, following the same setup as Cheng et al. (2023). As shown in Figure 7, updating V or DV significantly improves locality metrics, indicating the importance of aligning multimodal components. However, such updates lead to a drop of over 20% in relevance and generality, suggesting degraded model consistency. In contrast, editing LLM only with appropriate loss constraints yields a better balance between editing effectiveness and model retention. Note that when only the image encoder is updated, T_1I_4 and T_2I_4 reach 1 due to unchanged LLM parameters.

Layer	29-32	25-32	20-32	15-32	10-32	5-32	0-32
Model	Blip2OPT						
VQA	0.9506	0.5629	0.2382	0.2079	0.1773	0.0169	0.0017
VLKEB	0.8704	0.8065	0.6517	0.5218	0.3559	0.0473	0.0060
Model	MiniGPT4						
VQA	0.8966	0.8733	0.8233	0.7353	0.6453	0.5927	0.5820
VLKEB	0.8866	0.7740	0.6920	0.4707	0.2913	0.2267	0.1467

Table 1: Performance Change After Masking Unimportant Tokens. The results indicate that utilizing only the tokens we have pinpointed for prediction can maintain over 87% of the performance at higher layers.

Effect of Loss Functions. We investigate the impact of incorporating RI, NI, and CI losses individually and in combination (cf. Figure 7). The training data we use for each type of locality is: RI (T_1I_3), CI (T_2I_2), and NI (T_1I_4). Incorporating all three losses yields the best and most stable performance, achieving the highest scores in three out of four settings and ranking third on the Blip2OPT model with VQA. While using only RI+NI produces comparable results, it introduces instability on metrics such as T_1I_2 and T_1I_3, and performs well only on MiniGPT-4 with VLKEB. This instability arises from challenges in distinguishing edits involving the same image but differing attributions, for instance, identifying 'black shoes' versus 'yellow shoes' with the edits *The boy is wearing black shoes*. In such cases, more fine-grained constraints are necessary, which are effectively addressed by incorporating the CI loss. These results underscore the importance of jointly optimizing across diverse data types. By combining RI, NI, and CI losses, we enforce stronger consistency and robustness in editing performance across multimodal contexts. Thus we employ the CI-RI-NI as our loss combination.

Retrieval Consistency Analysis We also evaluate the consistency between text-based and image-based retrieval in multimodal model editing. We found that the average similarity between samples retrieved via image-based similarity and those retrieved via text-based encoding, as computed by CLIP, is 0.89. This result demonstrates a high degree of semantic consistency between the two retrieval methods, confirming that both text-based and image-based retrieval can yield comparably relevant samples.

6 Related Work

Model Editing. Model Editing has emerged as a viable strategy for precisely updating LLMs without the expensive resources (Wang et al. 2023; Zhang et al. 2024). A key

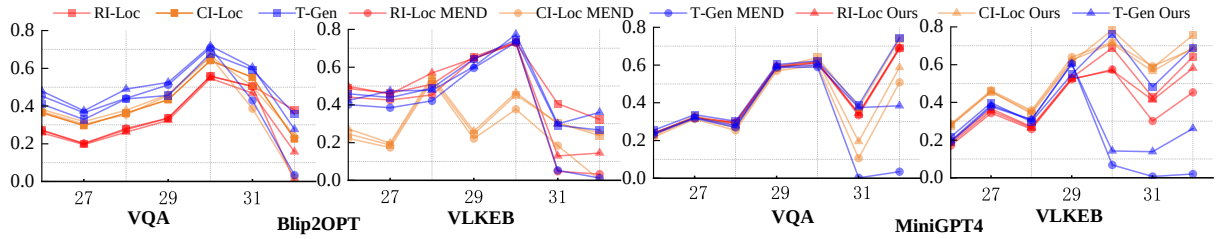

Figure 6: Contribution of image tokens in Blip2OPT and MiniGPT4, with evaluation across RI-Loc, CI-Loc, and T-Gen metrics. Squares, circles, and triangles denote results from the original model, MEND, and our method, respectively.

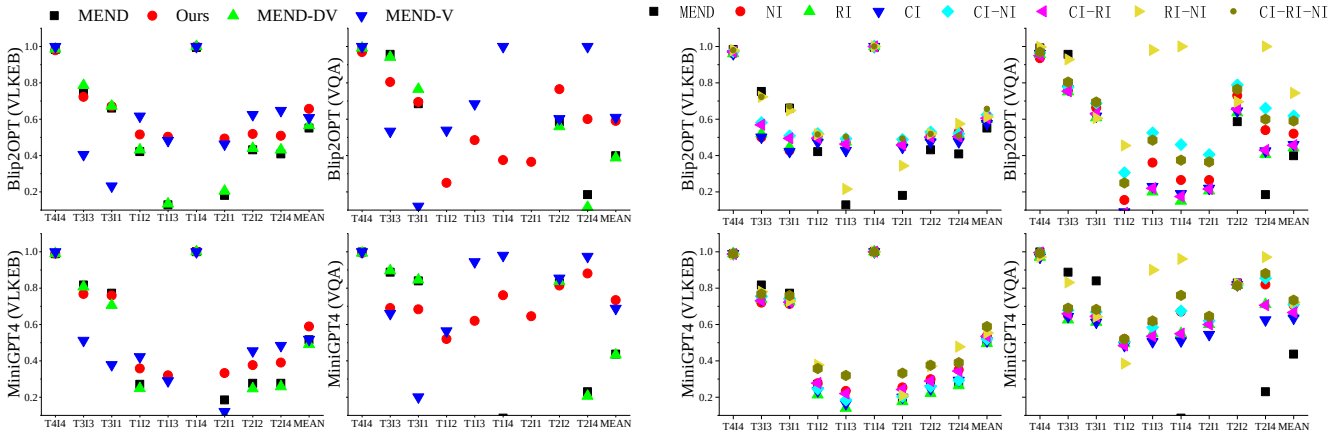

Figure 7: Ablation study results: (left) module updating comparison, where D refers to the last three layers of the LLM, V refers to the last three layers of the visual encoder, and DV is their combination; (right) loss function comparison.

challenge in model editing is how to update information without affecting unrelated data and without compromising the model's performance. (Hoelscher-Obermaier et al. 2023; Gupta, Rao, and Anumanchipalli 2024). Cheng et al. (2023) extend text-based editing methods to multimodal model editing and preliminarily verify its feasibility. However, the edited model only retains the predictions of the pre-edited model for randomly selected data. Compared to evaluating the success rate of updates, assessing the locality of editing models is more important for developing robust model editing methods. In this paper, we reveal transient blindness in MMED, as current evaluations overlook changes in multimodal abilities, resulting in inaccurate predictions for inputs that are related but differ from the edits, highlighting the important of Locality in MMED.

Editing in Multimodal Language Models. The development of large language models (LLMs) has spurred notable advances in multimodal LLMs (MLLMs) (Xu, Zhu, and Clifton 2023; Qi, Lv, and Ma 2025). These models typically leverage an LLM decoder to interpret fused image-text inputs, making the preservation of multimodal capabilities a central concern during model editing. Promising future solutions to this challenge include retrieval-augmented generation (Yan et al. 2025b,c) and memory-based mechanisms (Ma et al. 2025a; Wang et al. 2024).

7 Discussion & Conclusion

Our research uncovers critical limitations in current MMED evaluations, highlighting that existing metrics overlook cross-modal balance degradation and fail to detect transient blindness, a phenomenon where edited models over-rely on textual cues and disregard visual information. To address this, we propose **De-VQA**, a plug-and-play dynamic evaluation framework that introduces three comprehensive locality metrics and seven data types to assess multimodal editing effects. Through token attribution analysis, we trace transient blindness to imbalanced updates between textual and visual modalities, and our proposed locality-aware adversarial loss effectively balances cross-modal contributions. Experimental results demonstrate that our method outperforms existing baselines, improving locality preservation by 17% on average while maintaining high edit accuracy.

Our work not only revolutionises the evaluation of multimodal model editing by exposing transient blindness and refining locality assessment, but also offers a foundation for robust solutions to balance cross-modal updates, enabling more reliable knowledge correction in real-world vision-language applications, such as correcting misunderstandings in social media image-text posts (e.g., revising a satirical caption to prevent misinformation) and personalized content generation (e.g., learning a new corporate logo and its meaning for consistent marketing creation).

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (No.62376144) and the Natural Language Processing Innovation Team (Sanjin Talents) Project of Shanxi Province, the Science and Technology Cooperation and Exchange Special Project of Shanxi Province (No.202204041101016), the Key Research and Development Program of Shanxi Province (No.202102020101008), National Natural Science Foundation of China (No.62302297).

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Wang, J.; Qiao, Y.; Lin, D.; et al. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37: 27056–27087.

Chen, Q.; Wang, C.; Wang, D.; Zhang, T.; Li, W.; and He, X. 2025. Lifelong Knowledge Editing for Vision Language Models with Low-Rank Mixture-of-Experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 9455–9466.

Chen, Q.; Zhang, T.; He, X.; Li, D.; Wang, C.; Huang, L.; and Xue', H. 2024b. Lifelong Knowledge Editing for LLMs with Retrieval-Augmented Continuous Prompt Learning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13565–13580. Miami, Florida, USA: Association for Computational Linguistics.

Cheng, S.; Tian, B.; Liu, Q.; Chen, X.; Wang, Y.; Chen, H.; and Zhang, N. 2023. Can We Edit Multimodal Large Language Models? In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13877–13888. Singapore: Association for Computational Linguistics.

De Cao, N.; Aziz, W.; and Titov, I. 2021. Editing Factual Knowledge in Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6491–6506.

Du, Y.; Jiang, K.; Gao, Z.; Shi, C.; Zheng, Z.; Qi, S.; and Li, Q. 2025. MMKE-Bench: A Multimodal Editing Benchmark for Diverse Visual Knowledge. In *The Thirteenth International Conference on Learning Representations*.

Gupta, A.; Rao, A.; and Anumanchipalli, G. 2024. Model Editing at Scale leads to Gradual and Catastrophic Forgetting. arXiv:2401.07453.

Han, X.; Li, R.; Li, X.; Liang, J.; Zhang, Z.; and Pan, J. 2024. InstructEd: Soft-Instruction Tuning for Model Editing with Hops. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 14953–14968. Bangkok, Thailand: Association for Computational Linguistics.

Han, X.; Li, R.; Li, X.; and Pan, J. Z. 2023a. A divide and conquer framework for Knowledge Editing. *Knowledge-Based Systems*, 279: 110826.

Han, X.; Li, R.; Tan, H.; Yuanlong, W.; Chai, Q.; and Pan, J. Z. 2023b. Improving Sequential Model Editing with Fact Retrieval. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hoelscher-Obermaier, J.; Persson, J.; Kran, E.; Konstas, I.; and Barez, F. 2023. Detecting Edit Failures In Large Language Models: An Improved Specificity Benchmark. *arXiv preprint arXiv:2305.17553*.

Huang, H.; Zhong, H.; Yu, T.; Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2024. Vlkeb: A large vision-language model knowledge editing benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Huang, W.; Zhou, G.; Lapata, M.; Vougiouklis, P.; Montella, S.; and Pan, J. Z. 2025. Prompting Large Language Models with Knowledge Graphs for Question Answering involving Long-tail Facts. *Knowledge-Based Systems (KBS)*.

Huang, Z.; Shen, Y.; Zhang, X.; Zhou, J.; Rong, W.; and Xiong, Z. 2023. Transformer-Patcher: One Mistake Worth One Neuron. In *The Eleventh International Conference on Learning Representations*.

Jiang, Y.; Wang, Y.; Wu, C.; Zhong, W.; Zeng, X.; Gao, J.; Li, L.; Jiang, X.; Shang, L.; Tang, R.; Liu, Q.; and Wang, W. 2024. Learning to Edit: Aligning LLMs with Knowledge Editing. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4689–4705. Bangkok, Thailand: Association for Computational Linguistics.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Ma, B.; Li, R.; Yuanlong, W.; Tan, H.; and Li, X. 2025a. Memorization $\neq$ Understanding: Do Large Language Models Have the Ability of Scenario Cognition? In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 20758–20774. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Ma, Y.; Hong, X.; Zhang, S.; Li, H.; Zhu, Z.; Luo, W.; and Ma, Z. 2025b. ComprehendEdit: A Comprehensive Dataset and Evaluation Framework for Multimodal Knowledge Editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19323–19331.

Mitchell, E.; Lin, C.; Bosselut, A.; Finn, C.; and Manning, C. D. 2022a. Fast Model Editing at Scale. In *International Conference on Learning Representations*.

Mitchell, E.; Lin, C.; Bosselut, A.; Manning, C. D.; and Finn, C. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, 15817–15831. PMLR.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pan, J. Z.; Razniewski, S.; Kalo, J.-C.; Singhania, S.; Chen, J.; Dietze, S.; Jabeen, H.; Omeliyanenko, J.; Zhang, W.; Lissandrini, M.; Biswas, R.; de Melo, G.; Bonifati, A.; Vakaj, E.; Dragoni, M.; and Graux, D. 2023. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1–38.

Pi, R.; Han, T.; Xiong, W.; Zhang, J.; Liu, R.; Pan, R.; and Zhang, T. 2025. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, 382–398. Springer.

Qi, M.; Lv, C.; and Ma, H. 2025. Robust Disentangled Counterfactual Learning for Physical Audiovisual Commonsense Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Sinitsin, A.; Plokhotnyuk, V.; Pyrkin, D.; Popov, S.; and Babenko, A. 2019. Editable Neural Networks. In *International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Kai, S.; Luo, L.; Wei, W.; Hu, Y.; Liew, A. W.-C.; Pan, S.; and Yin, B. 2024. Large language models-guided dynamic adaptation for temporal knowledge graph reasoning. *Advances in Neural Information Processing Systems*, 37: 8384–8410.

Wang, R.; and Li, P. 2024. LEMoE: Advanced Mixture of Experts Adaptor for Lifelong Model Editing of Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2551–2575. Miami, Florida, USA: Association for Computational Linguistics.

Wang, S.; Zhu, Y.; Liu, H.; Zheng, Z.; Chen, C.; et al. 2023. Knowledge Editing for Large Language Models: A Survey. *arXiv preprint arXiv:2310.16218*.

Wu, K.; Jiang, B.; Jiang, Z.; He, Q.; Luo, D.; Wang, S.; Liu, Q.; and Wang, C. 2024. NoiseBoost: Alleviating Hallucination with Noise Perturbation for Multimodal Large Language Models. *arXiv preprint arXiv:2405.20081*.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yan, Z.; Wang, J.; Chen, J.; Li, X.; Li, R.; and Pan, J. Z. 2025a. Decomposing and Revising What Language Models Generate. In *In the Proceedings of the 28th European Conference on Artificial Intelligence (ECAI-2025)*.

Yan, Z.; Wang, J.; Chen, J.; Li, X.; Liang, J.; Li, R.; and Pan, J. Z. 2025b. Atomic fact decomposition helps attributed question answering. *IEEE Transactions on Knowledge and Data Engineering*.

Yan, Z.; Wang, J.; Chen, J.; Wang, Y.; Tan, H.; Liang, J.; Li, X.; Li, R.; and Pan, J. Z. 2025c. Prompting large language models with partial knowledge for answering questions with unseen entities. *arXiv preprint arXiv:2508.01290*.

Zhang, N.; Yao, Y.; Tian, B.; Wang, P.; Deng, S.; Wang, M.; Xi, Z.; Mao, S.; Zhang, J.; Ni, Y.; et al. 2024. A Comprehensive Study of Knowledge Editing for Large Language Models. *arXiv preprint arXiv:2401.01286*.

Zheng, C.; Li, L.; Dong, Q.; Fan, Y.; Wu, Z.; Xu, J.; and Chang, B. 2023. Can We Edit Factual Knowledge by In-Context Learning? In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4862–4876. Singapore: Association for Computational Linguistics.

Zheng, D.; Lapata, M.; and Pan, J. Z. 2025. Long-Form Information Alignment Evaluation Beyond Atomic Facts. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 11018–11038. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Zhu, C.; Rawat, A. S.; Zaheer, M.; Bhojanapalli, S.; Li, D.; Yu, F.; and Kumar, S. 2020. Modifying Memories in Transformer Models. arXiv:2012.00363.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*.