Towards Robust and Scalable Knowledge Editing in Text-to-Image Diffusion Models

Yifei Liu YIFEIL25@MAILS.JLU.EDU.CN
Xin Wang* XINWANG@JLU.EDU.CN
School of Artificial Intelligence, Jilin University

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Knowledge editing in Text-to-Image(T2I) diffusion models aims to update specific factual associations without disrupting unrelated knowledge. However, existing methods often suffer from unintended collateral effects, where editing a single fact can alter the representation of non-target named entities, degrading generation quality for unrelated prompts, which becomes more severe in real-world, dynamic environments requiring frequent updates. To address this challenge, we introduce a novel editing framework supporting large-scale T2I knowledge editing. Our framework incorporates our proposed Entity-Aware Text Alignment(EATA) to penalize unintended changes in unaffected entities and employs a principled null-space projection strategy to minimize perturbations to existing knowledge. Experimental results demonstrate that our approach enables precise and robust large-scale T2I knowledge editing, preserves the integrity of unrelated content, and maintains high generation fidelity, while offering scalability for continuous editing scenarios.

Keywords: Text-to-Image Diffusion Models, T2I Knowledge Editing, Prompt-based Generation

1. Introduction

Text-to-Image (T2I) generative models Ho et al. (2020); Ramesh et al. (2022); Rombach et al. (2022); Ho et al. (2022); Croitoru et al. (2023); Cao et al. (2024) have rapidly advanced in recent years, enabling the synthesis of photorealistic images from natural language descriptions. Powered by large-scale diffusion architectures and trained on massive datasets of image-text pairs, these models have internalized a wide range of factual and commonsense knowledge. However, due to the static nature of their training data, the information embedded in these models is inherently fixed at the time of training. As real-world facts evolve, these models may generate outdated or incorrect content, limiting their reliability in dynamic or time-sensitive applications. Retraining such large models to reflect updated knowledge is computationally intensive, time-consuming, and often impractical. Therefore, there is a pressing need for fast, low-cost techniques that enable efficient and precise updates to a model's internal knowledge without full retraining or reliance on prompt engineering.

To address this problem, most prior studies have primarily focused on model editing methods. These approaches typically involve fine-tuning or precisely targeting specific layers to make minimal yet effective modifications to the model's distribution. This allows for the desired concept to be updated while preserving the original knowledge. Existing

* Corresponding author.

methods often concentrate on editing either the cross-attention layers or the multi-layer perceptron (MLP) layers to update new concepts. They generally achieve this by aligning the representation of a source prompt (e.g., "The president of the United States") with that of a target prompt (e.g., "Tim Cook").

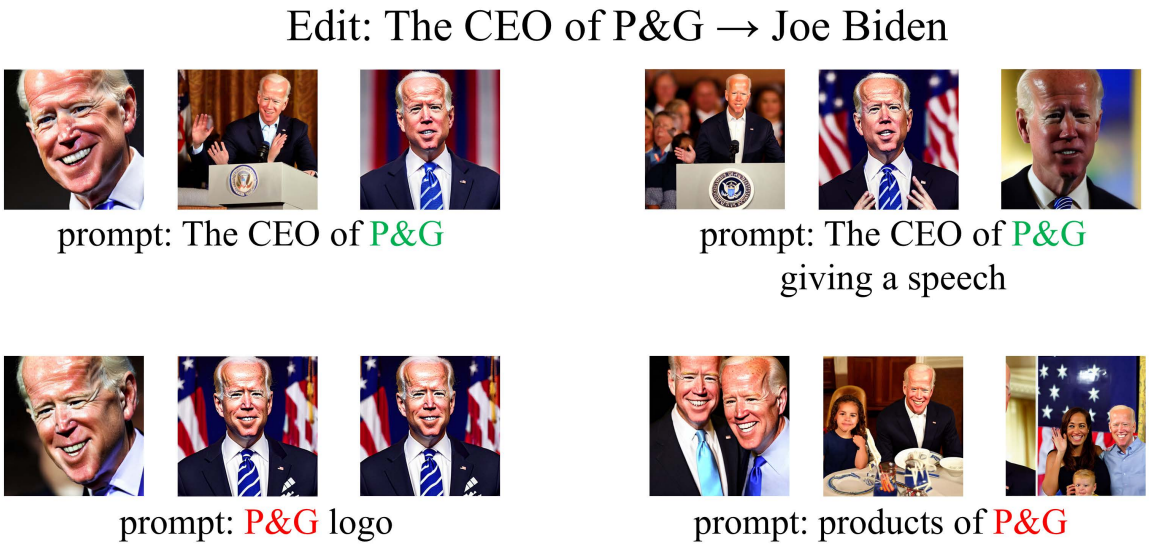

Figure 1: Successful and failed generated results of EMCID Xiong et al. (2024) when conducting the edit "The CEO of P & G → Joe Biden". The results demonstrate how the named entity within the prompt text are affected during the alignment of the source and target prompts.

However, despite the effectiveness of existing T2I knowledge editing methods, several important issues remain to be addressed. First, few studies have explored large-scale knowledge editing in Text-to-Image models. Existing research on large-scale T2I knowledge editing for diffusion models is still limited, and their performance leaves substantial room for improvement across multiple dimensions. Second, current methods rarely consider fine-grained alignment of sentence-level semantic representations. In particular, they often overlook the effects on subword or token-level semantics within the source prompt, which can undermine the fidelity of the edited model. For instance, as shown in Fig. 1, when aligning the source prompt "The CEO of P & G" with the target prompt "Joe Biden", the model may successfully generate images of Joe Biden in response to the edited prompt. However, it fails to produce correct outputs for prompts like "P & G logo" or "products of P & G", which contain the entity "P & G" but are unrelated to the edited concept. This suggests that the token-level representation of "P & G" was unintentionally affected during the editing process, leading to semantic misalignment and incorrect generations.

These observations highlight the limitations of current approaches and underscore the need for a new method capable of precisely editing specific concepts while preserving the semantics of unrelated content.

In summary, this paper's primary contributions are encapsulated as follows:

- We introduce a novel framework, which enables large-scale T2I knowledge editing while preserving internal knowledge and maintaining the overall performance of the diffusion models.
- We propose EATA (Entity-Aware Text Alignment) Loss combined with sentence level text alignment loss, which is able to effectively protect the model's understanding of named entities while facilitating successful T2I knowledge editing.
- We adopt a closed-form solution with null-space projection, which removes the need for manually balancing old and new knowledge in the optimization objective.
- We compare our proposed method with existing T2I knowledge editing approaches. Through both quantitative and qualitative analyses, our method demonstrates superior effectiveness in performing large-scale T2I knowledge editing while preserving the overall performance of the model.

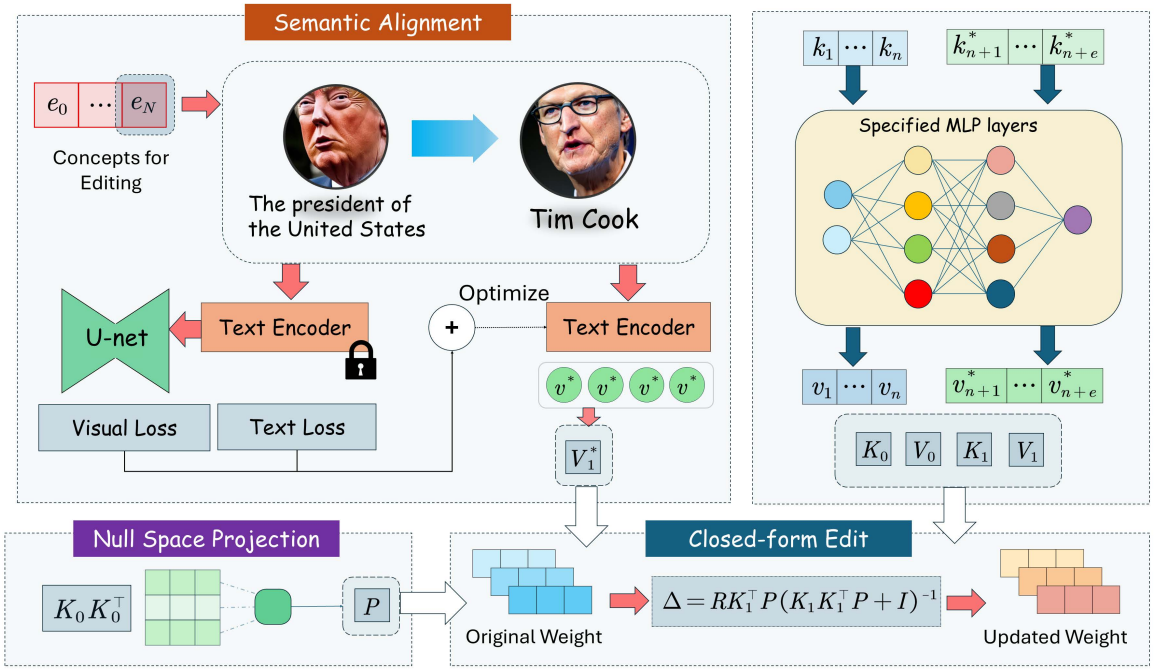

Figure 2: Overview of our proposed framework, which consists of three main stages. **Null Space Projection:** a projection matrix is estimated using the covariance of the original concept keys K_0. **Semantic Alignment:** source and target prompts are aligned via a combination of text alignment and visual alignment losses, yielding refined value vectors $V_1 \rightarrow V_1^*$. **Closed-form Edit:** the final weight update is performed by applying (K_1, V_1^*) to the closed-form solution and adding perturbation to the original weight matrix.

2. Related Work

Recent years have witnessed growing interest in editing diffusion-based T2I diffusion models to incorporate new knowledge . A variety of approaches Gandikota et al. (2023); Zhang et al. (2024); Fan et al. (2023); Kumari et al. (2023); Zheng and Yeh (2024); Heng and Soh (2023) have been explored to achieve this goal through fine-tuning techniques. However, continual fine-tuning tends to degrade the overall performance of the model, leading to the well-known catastrophic forgetting phenomenon McCloskey and Cohen (1989); Kirkpatrick et al. (2017). This makes it unsuitable as a long-term solution for concept editing. Moreover, frequent fine-tuning incurs high computational costs Mitchell et al. (2022), limiting its practical applicability.

Inspired by earlier work on model editing Meng et al. (2022a,b), several approaches have explored model editing methods using closed-form solutions. TIME Orgad et al. (2023) edits implicit assumptions in diffusion models by adjusting projection matrices in cross-attention layers, aligning source prompt (e.g., "a pack of roses") with destination prompt (e.g., "a pack of blue roses"). ReFACT Arad et al. (2023) focuses on the text encoder's MLP layers in CLIP Radford et al. (2021); Ilharco et al. (2021), treating them as linear associative memories to update key-value mappings via closed-form solutions. Several approaches Gandikota et al. (2023); Gong et al. (2024); Lu et al. (2024); Xiong et al. (2024); Wu and Harandi (2024) utilize closed-form editing to perform concept erasure, demonstrating impressive effectiveness in eliminating specific knowledge. Although ReFACT and TIME perform well for single-concept editing, they are not designed to scale to large-scale editing scenarios. In contrast, our method is capable of handling large-scale knowledge updates, while ensuring both faithful concept injection and high-quality image generation across diverse prompts.

For large-scale editing, EMCID Xiong et al. (2024) introduces a two-stage framework that supports multiple edits on diffusion models, maintaining high edit fidelity while alleviating catastrophic forgetting. However, EMCID requires a delicate trade-off between preserving prior knowledge and integrating new information, and it becomes especially problematic when editing concepts involving named entities, thereby causing noticeable degradation in output quality. In this research, our proposed framework eliminates the need to balance old and new knowledge in the overall objective. Also, we introduce a more fine-grained text alignment mechanism. This design leads to more precise factual updates, better preservation of existing knowledge, and improved overall generation quality—especially under large-scale, dynamic editing scenarios.

3. Methods

3.1. Preliminaries

Text-to-Image Diffusion Models Text-to-image (T2I) diffusion models are typically based on denoising diffusion probabilistic models (DDPMs) Ho et al. (2020), which learn to reverse a gradual noising process through iterative denoising steps. Given a text prompt y, the model is trained to generate an image x by sampling from the conditional distribution $p_\theta(x|y)$. The training objective is to learn a denoising network ϵ_θ that predicts the noise added at each timestep t in the forward diffusion process.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}) \tag{1}$$

Here, x_t denotes the noisy image at timestep t, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and y is the conditioning text input, typically embedded via a pretrained language model. During generation, the model samples $x_T \sim \mathcal{N}(0, \mathbf{I})$ and applies the learned reverse process to iteratively denoise and recover x_0.

Model Editing with Closed-form Solution Several existing editing methods, such as ROME and MEMIT, have demonstrated strong performance in T2I knowledge editing tasks on autoregressive large language models (LLMs) Radford et al. (2019); Brown et al. (2020); Devlin et al. (2019). Through the causal tracing of factual associations, ROME Meng et al. (2022a) reveals that feedforward MLPs at a range of middle layers are decisive when processing the last token of the subject name. Also, ROME treats the transformer MLP as a linear associative memory. By solving $\hat{W} = W + \Lambda(C^{-1}k_*)^T$, where $\Lambda = (v_* - Wk_*)/((C^{-1}k_*)^T k_*)$, and $C = KK^T$ is the precomputed covariance of input keys, ROME inserts a new key-value pair (k_*, v_*) to update factual associations after perturbing W, ensuring minimal interference with existing memories. Building on this, MEMIT Meng et al. (2022b) proposes a scalable multi-layer method that distributes updates across critical MLP layers, which stack the k and v from the preserved and new knowledge as K_0/V_0 and K_1/V_1 respectively. By solving $\Delta = RK_1^T(K_0K_0^T + K_1K_1^T)^{-1}$, where $R = V_1 - W_0K_1$ represents residual errors for new memories. This approach enables bulk editing of thousands of facts while maintaining the performance of the model.

3.2. The Overview of the Framework

Our proposed framework, as illustrated in Fig. 2, consists of three key stages, that is, null space projection, semantic alignment and closed-form edit. Firstly, the initialization of source prompts p_{src} (e.g., "The president of the United States") and target prompts p_{tgt} (e.g., "Tim Cook") is conducted, which can derive the key-value pairs of the original and new concepts, respectively. As demonstrated in previous studies Kohonen (2009); Meng et al. (2022a), the multi-layer perceptron (MLP) layers in the text encoder contain two matrices separated by a non-linear activation function. This structure can be formulated as $W_{proj} \cdot \sigma(W_{fc})$, thereby constructing a linear projection for the key-value stores, i.e., $WK \approx V$, while $(K_0|V_0)$ and $(K_1|V_1)$ correspond to the key-value pairs of original and new concepts, respectively. Specifically, when providing $K_0 = [k_1|\cdots|k_n]$ and $K_1 = [k_{n+1}|\cdots|k_{n+e}]$, the linear memory association yields the corresponding value vectors $V_0 = [v_1|\cdots|v_n]$ and $V_1 = [v_{n+1}|\cdots|v_{n+e}]$. In the null space projection stage, we compute the projection matrix P based on the covariance of K_0. Subsequently, both source and target prompts are utilized for semantic alignment. The optimization is guided by a combination of text alignment loss containing our proposed non-target token loss, and noise prediction loss, the target value vectors V_1 are refined into V_1^*. All outputs from the preceding steps contribute to a closed-form edit, which updates the weight matrices of the text encoder and ultimately completes the T2I knowledge editing process. Based on previous studies Meng et al. (2022a); Arad et al. (2023), we identify the text encoder within the diffusion model as the key component for performing editing.

3.3. Null Space-Guided T2I Knowledge Editing

In the context of large-scale concept editing, existing studies Meng et al. (2022b); Xiong et al. (2024) primarily focus on achieving a trade-off between preserving the original concept and incorporating the new ones, which is typically formulated as:

$$W^* = \arg\min_W \left(\|WK_1 - V_1\| + \gamma\|WK_0 - V_0\|\right) \tag{2}$$

However, this optimization objective relies on the balancing parameter γ, which makes it a struggle to contrive a trade-off between retaining existing knowledge and integrating newly edited information.

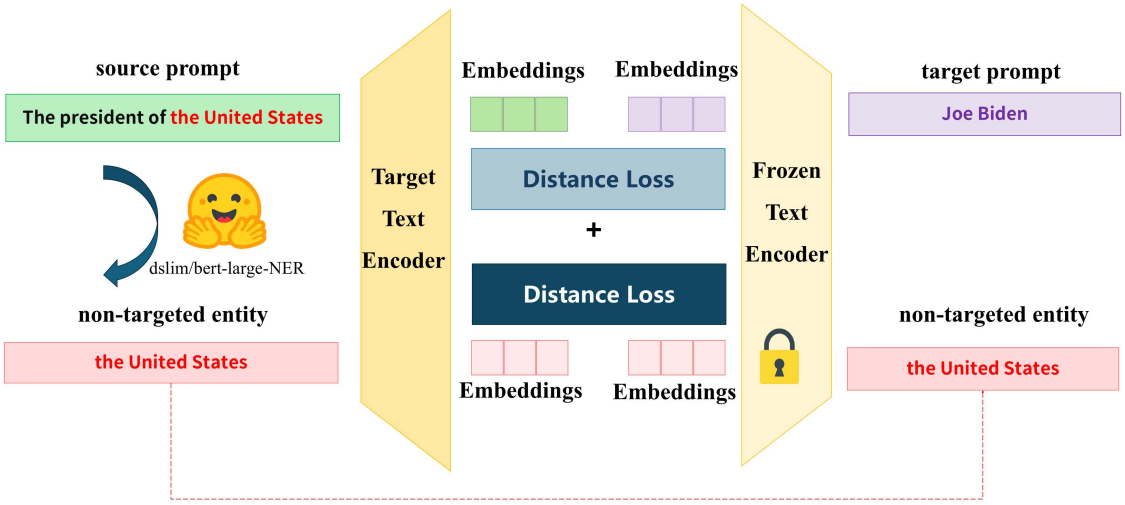

Figure 3: Illustration of the proposed EATA method. We first utilize the Named Entity Recognition (NER) model to extract potential named entities from the source prompt. Then, representations are obtained from both a frozen text encoder and a target text encoder. Two types of losses are computed: the general sentence-level text alignment loss and our proposed EATA loss. These two losses are then combined to form the overall text alignment loss used for optimization.

Inspired by the study Fang et al. (2024), we adopt a null space projection approach to address this dilemma. By computing the projection matrix onto the null space of the covariance matrix $K_0K_0^T$, we obtain a projection matrix P, which is then combined with the perturbation Δ of the original weight matrix W. After applying this transformation, $(W+\Delta)K_0$ can be treated as equivalent to WK_0, allowing us to focus solely on minimizing the distance loss for the new knowledge, as shown below:

$$W^* = \arg\min_W \|WK_1 - V_1\| \tag{3}$$

Source Prompt	Extracted Entity Tokens
The president of the United States	the United States
The CEO of Amazon	Amazon
The chief scientist at NASA	NASA
The lead singer of Beatles	Beatles

Table 1: Examples of extracted entities from source prompts. The extracted tokens of entities will be used for calculating the Entity-Aware Text Alignment Loss.

3.4. Entity-Aware Text Alignment

To ensure that the generated images align with expectations, the value vectors V_1 must be refined so that they yield correct results when the context includes the source prompt. Therefore, it is necessary to optimize and update the refined vectors v^* within V_1.

Considering that T2I diffusion models are conditioned on a text prompt that guides the image generation process. it's crucial to align the representation of the text prompt. Here we introduce the **Entity-Aware Text Alignment (EATA)**, which applies a named entity recognition (NER) module to extract salient proper nouns or entities from the source prompt (Examples shown in Tab. 1). By incorporating the prompt text alignment loss and EATA loss, we then are able to separate alignment constraints not only between p_{src} and p_{tgt} (e.g., "The president of the United States" $\rightarrow$ "Tim Cook"), but also between non-target entities and themselves (e.g., "the United States" $\rightarrow$ "the United States"). Here we got the formulation of the loss function like:

$$\mathcal{L}_{EATA} = \|c_{src} - c_{tgt}\|^2 + \alpha\|c_e - c_e\|^2 \tag{4}$$

where c_{src} and c_{tgt} denote the representations, specifically the feature vector of the last subject token [EOS] from the source and target prompts derived from the text encoder, and c_e represents the named entity extracted from the source prompt. To facilitate entity extraction, we employ the bert-large-NER Tjong Kim Sang and De Meulder (2003), a fine-tuned BERT model that achieves state-of-the-art performance on named entity recognition tasks. The coefficient α, which balances the two loss components, is empirically set to 0.11.

Additionally, to further enhance the quality of the generated images in terms of both semantic accuracy and visual fidelity, we incorporate a visual alignment loss. This loss is applied during the noise prediction stage, encouraging consistency between the edited and target prompts in the denoising process. The loss is formulated as follows:

$$\mathcal{L}_V = \|\epsilon(\mathbf{x}_t, c_{src}, t) - \epsilon(\mathbf{x}_t, c_{tgt}, t)\|^2 \tag{5}$$

where $\epsilon(\mathbf{x}_t, c_{src}, t)$ denotes the predicted noise conditioned on the source prompt, and $\epsilon(\mathbf{x}_t, c_{tgt}, t)$ denotes the predicted noise conditioned on the target prompt.

We combine this with our entity-aware text alignment loss to form the overall optimization objective for each concept editing instance:

$$\mathcal{L} = \mathcal{L}_{\text{EATA}} + \lambda\mathcal{L}_V \tag{6}$$

Here, λ is a balance factor that controls the contribution of the visual alignment loss. Following the practice in Xiong et al. (2024), we empirically set $\lambda = 0.01$.

3.5. Closed-form Edit

As for the closed-form edit, this stage utilizes the previously derived value to form the perturbation of the original weight matrix within the multiple layers of text encoder. The closed-form solution is displayed as below:

$$\Delta = RK_1^T P(K_1 K_1^T P + I)^{-1} \tag{7}$$

We define the residual vector of the current edit as $R = V_1 - WK_1$, where P denotes the null space projection matrix obtained in the previous step. The identity matrix I ensures invertibility and improves numerical stability, turning a possibly singular semidefinite matrix into a positive definite one Fang et al. (2024). We then compute the perturbation Δ and add it to the original weight matrix W to obtain the updated weight matrix in the text encoder.

4. Experiment

4.1. Experiment Setup

In our experiments, we utilize Stable Diffusion v1.4 (CompVis/stable-diffusion-v1-4) Rombach et al. (2022) as the target model. It is a latent T2I diffusion model capable of generating high-quality, photorealistic images from diverse natural language prompts. All experiments are conducted using an NVIDIA RTX 4090 GPU. To measure the semantic alignment between the generated images and the input prompts, we employ CLIP ViT-bigG/14 (trained on LAION-2B) Schuhmann et al. (2022).

Dataset We evaluate our method on three benchmark datasets for T2I diffusion model editing. The TIMED dataset Orgad et al. (2023) focuses on attribute-level modifications, where each entry includes a source prompt (e.g., "a dog") and a destination prompt (e.g., "a green dog") to test whether the model can integrate new attributes while preserving original semantics. It also provides five positive and five negative prompts to assess generalization and retention. The RoAD dataset Arad et al. (2023) extends TIME by emphasizing role-based edits involving entities like politicians or musicians. Each of its 100 entries includes an edit prompt, a source prompt, a target prompt, and the same structure of positive/negative prompts. The CAKE dataset Gu et al. (2024) offers a more comprehensive evaluation, introducing multi-entity scenes and complex semantics. It contains 100 editing entries and 1,500 evaluation prompts, categorized into Edit I (single-object) and Edit II (multi-object), with detailed subcategories such as Efficacy, Generality, Specificity, KgeMap, and Compo for fine-grained behavioral analysis.

Baseline Given the large-scale T2I knowledge editing scenario, we adopt three representative methods as our baselines: TIME Orgad et al. (2023), ReFACT Arad et al. (2023), and EMCID Xiong et al. (2024). For a clearer comparison, we also include an Oracle model—i.e., the original model without any edits for comparison. Due to differences in dataset formats and evaluation protocols, we follow the approach of Gu et al. (2024) and adapt the datasets accordingly to ensure compatibility with all baseline methods.

Dataset	Method	Score	Efficacy	Generality	Specificity	FID(↓)	CLIP
TIMED	Oracle	49.17	25.00%±2.65	50.23%±1.89	**94.65**%±1.78	33.41	0.465
	TIME	0.00	0.00%±0.00	0.00%±0.00	1.50%±0.39	34.21	0.464
	ReFACT	43.49	22.50%±3.31	41.58%±1.01	87.92%±2.09	33.84	0.465
	EMCID	65.07	70.19%±2.19	64.42%±1.51	60.92%±3.64	33.80	0.465
	Ours	**71.60**	**73.85**%±4.23	**69.15**%±2.91	71.88%±2.21	33.66	0.465
RoAD	Oracle	16.59	3.56%±1.30	13.33%±0.37	**96.36**%±1.50	33.41	0.465
	TIME	0.85	0.44%±0.54	0.27%±0.00	5.21%±0.42	33.54	0.464
	ReFACT	15.07	2.89%±0.89	12.89%±1.15	91.96%±1.81	34.36	0.465
	EMCID	62.72	69.11%±2.15	51.78%±1.55	68.93%±1.28	33.79	0.465
	Ours	**70.14**	**77.56**%±2.37	**62.27**%±2.90	71.47%±0.91	33.58	0.465

Table 2: Evaluation results on **TIMED** and **RoAD** datasets. Best results are marked with **bold**, and best among editing methods are marked with underline.

Method	Score	Efficacy	Generality	KgeMap	Compo	Specificity	FID (↓)	CLIP
Oracle	0.00	00.00%±0.00	02.72%±0.96	03.27%±0.65	01.73%±0.48	**96.06**%±1.57	33.41	0.465
TIME	0.00	00.00%±0.00	00.00%±0.00	00.00%±0.00	00.00%±0.00	01.27%±0.44	33.68	0.465
ReFACT	0.00	00.00%±0.00	02.92%±0.82	03.20%±0.72	01.73%±0.53	91.40%±1.12	34.17	0.465
EMCID	24.52	61.40%±5.04	26.84%±3.06	22.80%±2.86	13.00%±3.41	18.13%±2.35	33.93	0.464
Ours	40.06	**63.40**%±5.24	**43.12**%±7.18	**34.67**%±3.77	**28.60**%±5.74	38.07%±3.53	33.81	0.465

Table 3: Evaluation results on the dataset CAKE. Best results are marked with **bold**. Best among editing methods are marked with underline. **Score** refers to the geometric mean of all the five metrics, **FID** refers to FID-5K, **CLIP** refers to the average CLIP score.

Metrics To comprehensively evaluate the effectiveness of T2I knowledge editing methods, we adopt a suite of metrics that capture different aspects of model behavior post-editing. **Efficacy** measures how accurately the model generates images that reflect the edited concept when given the original editing prompt. **Generality** assesses the model's ability to generalize the edited concept to semantically related prompts in varied contexts. **Specificity**, on the other hand, evaluates whether the editing process affects unrelated concepts, ensuring that the model retains its original behavior on prompts outside the scope of the edit. Additionally, we incorporate two metrics specified in CAKE dataset which is proposed by Gu et al. (2024). **KgeMap**, which examines whether the model can correctly respond to paraphrased or semantically similar prompts of the target concept, and **Compo**, which evaluates the model's ability to coherently combine and represent multiple edited concepts within a single prompt. Together, these metrics provide a comprehensive assessment of the edit's precision, generalization, and unintended side effects. we also compute the FID-5k and CLIP on the MS-COCO validation dataset Lin et al. (2014).

4.2. Quantitive Evaluation

To assess model's ability to edit multiple facts, we perform the large-scale concept edit, that is, update all the concepts from the items in the dataset at a time. While in the semantic alignment stage, we set the number of gradient steps as 200 with learning rate determined as the value 0.2. We use 5 random seeds, editing a clean model and generating one image per prompt for each seed. We then compute each of the metrics using CLIP as a zero-shot classifier.

Based on the results obtained from large-scale T2I knowledge editing experiments, shown as Tab. 2 and Tab. 3, we observe that while ReFACT and TIME perform reasonably well on the TIMED and RoAD datasets, showing satisfactory post-edit generation quality, yet their performance on the CAKE dataset degrades significantly. In particular, both methods suffer a substantial drop in effectiveness, which suggests a potential occurrence of catastrophic forgetting. This indicates that although ReFACT and TIME aim to preserve the model's original knowledge during editing, they struggle to maintain the balance between editing precision and the retention of prior knowledge. As a result, their edited generations often fail to meet the actual needs of users.

In contrast, after large-scale editing, our method can maintain the model's fundamental performance. Compared to other approaches, it not only more accurately reflects the intended edits but also better preserves the core capabilities of the model. Moreover, our method exhibits superior generalization and strong adaptability to contextual text.

4.3. Qualitative Evaluation

In order to further evaluate the effectiveness of our T2I knowledge editing method, we perform concept editing and image generation on three datasets: TIMED, RoAD, and CAKE. As shown in the Fig. 4, our method successfully accomplishes the editing tasks while maintaining high levels of fidelity and image quality in the generated outputs. These results demonstrate that our approach is effective for large-scale T2I knowledge editing and can robustly update the model's internal representations without compromising generation performance.

In addition, we compare our framework with existing methods by generating images conditioned on Efficacy and Specificity prompts. As shown in Fig. 5, the generations produced by diffusion models edited with TIME and ReFACT reveal several limitations in T2I knowledge editing, particularly in handling proper nouns that are unrelated to the original content, and exhibit varying levels of interference with the model's underlying knowledge. Specifically, TIME fails to generate coherent outputs for the edited concepts after large-scale modifications. While ReFACT is able to produce fluent text, it does not accurately reflect the intended edits. EMCID successfully incorporates the target edits even under large-scale changes; however, it tends to generate irrelevant content when prompted with unrelated concepts, indicating poor specificity.

In contrast, our method not only preserves generation quality and successfully incorporates the intended edits but also maintains robustness in handling prompts involving unrelated proper nouns, thus achieving a better trade-off between fidelity and specificity.

Figure 4: The generation results selected after the T2I knowledge editing on three datasets. The results demonstrate the excellent performance of our framework in large-scale T2I knowledge editing tasks.

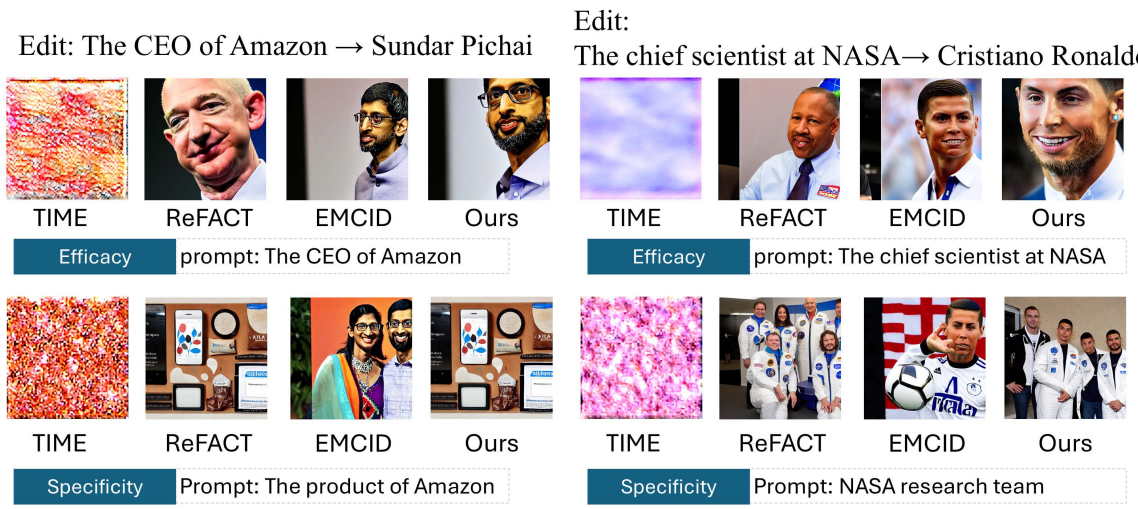

Figure 5: Visual comparison of image quality produced by different T2I knowledge editing algorithms.

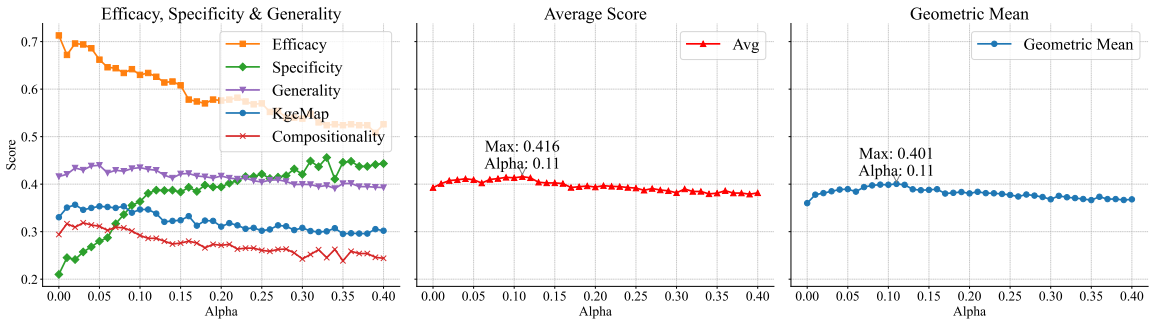

Figure 6: The evaluation results under different trade-off parameters α, where Average Score denotes the arithmetic mean and Geometric Mean refers to the geometric mean.

4.4. Empirical Analysis of Loss Weighting Strategies

To explore the optimal parameter that balances the trade-off between the text-alignment loss and the EATA loss, we perform a systematic grid search over the weight coefficient α with a step size of 0.01. For each value of α, we carry out a complete round of large-scale T2I knowledge editing based on dataset CAKE, and evaluate the model's performance using multiple metrics, including **Specificity**, **Efficacy**, **Generality**, **KgeMap** and **Compo**. In addition to individual metrics, we also compute the **arithmetic mean** and **geometric mean** of all the metrics calculated, which offer a more comprehensive assessment of performance across competing objectives. The result is shown as Fig. 6. As demonstrated in the first subfigure, there exists a subtle trade-off between Efficacy and Specificity that enhancing the preservation of unrelated concepts often comes at the cost of reduced precision in modifying the target concept.

From the results displayed in Fig. 6, the model's behavior varies notably with different values of α. By analyzing both the arithmetic and geometric means across metrics, we identify $\alpha = \alpha^*$ (set to 0.11) as the optimal value. Consequently, we adopt this setting for all experiments using our proposed framework. Our experimental results show that when the EATA loss is not introduced (i.e., $\alpha = 0$), the overall geometric and arithmetic mean scores are 34.8% and 38.82%, respectively. In contrast, after tuning the α parameter, the geometric mean can improve to 40.1% and the arithmetic mean to 41.6%, which validates the effectiveness of our method in improving generation performance.

5. Conclusion

In this paper, we propose a novel framework that not only supports large-scale T2I knowledge editing, but preserves existing knowledge in the model while minimizing performance degradation. We employ a more fine-grained alignment at the token level, particularly focusing on named entities, to ensure these concepts remain unaffected and intact. Additionally, we incorporate a null-space projection technique to restrict the proportion of affected concepts within the model. Experimental results demonstrate that our method can successfully perform large-scale T2I knowledge editing while preserving the overall per-

formance of the model, with minimal impact on non-target knowledge. Our work offers a new perspective on model editing, paving the way for future research into more robust and reliable T2I knowledge editing techniques.

6. Limitations

Although our proposed method demonstrates strong performance in large-scale knowledge editing tasks for diffusion models and effectively maintains model performance after editing, it occasionally generates images with unnatural or slightly distorted effects, shown as Fig 7 (a). In addition, for certain entities (e.g., "The author of 1984"), entity extraction may fail, which in turn affects the quality of concept editing and manifests in the image results generated from specificity-based prompts, shown as Fig 7 (b). This suggests the need for further investigation into methods that stabilize the quality of image generation and for developing more robust and adaptable non-target entity text alignment approaches.

7. Ethical Statement

Our method preserves the model's original knowledge while enabling large-scale concept editing. However, like other T2I diffusion models, it cannot fully prevent inappropriate or NSFW outputs under low-toxicity or ambiguous prompts—a known limitation of diffusion models. All prompts in our experiments are used solely for academic evaluation, without intent to offend or misrepresent any individuals or organizations. We advocate responsible and ethical use of generative models.

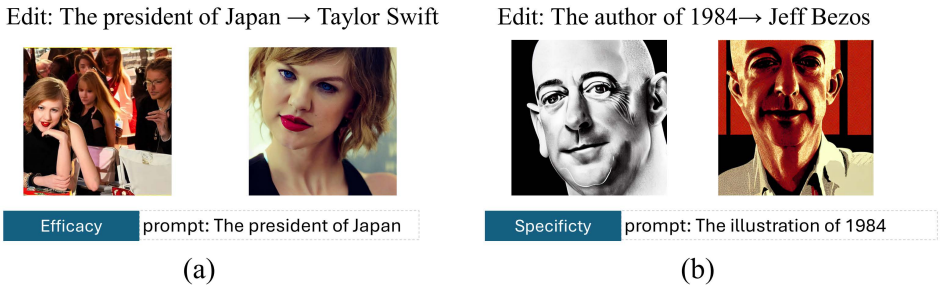

Figure 7: Failed cases of our method. (a) illustrates that occasional distortions or unnatural artifacts may occur in the generated images, while (b) shows that failures in entity extraction can lead to suboptimal editing results.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China under grants (No.62372211, 62272191), and the Science and Technology Development Program of Jilin Province (No.20250102216JC).

References

Dana Arad, Hadas Orgad, and Yonatan Belinkov. Refact: Updating text-to-image models by editing the text encoder. *arXiv preprint arXiv:2306.00738*, 2023.

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

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

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.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pages 73–88. Springer, 2024.

Hengrui Gu, Kaixiong Zhou, Yili Wang, Ruobing Wang, and Xin Wang. Pioneering reliable assessment in text-to-image knowledge editing: Leveraging a fine-grained dataset and an innovative criterion. *arXiv preprint arXiv:2409.17928*, 2024.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36: 17170–17194, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Teuvo Kohonen. Correlation matrix memories. *IEEE transactions on computers*, 100(4): 353–359, 2009.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

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, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

Jing Wu and Mehrtash Harandi. Munba: Machine unlearning via nash bargaining. *arXiv preprint arXiv:2411.15537*, 2024.

Tianwei Xiong, Yue Wu, Enze Xie, Zhenguo Li, and Xihui Liu. Editing massive concepts in text-to-image diffusion models. *arXiv preprint arXiv:2403.13807*, 2024.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1755–1764, 2024.

Amber Yijia Zheng and Raymond A Yeh. Imma: Immunizing text-to-image models against malicious adaptation. In *European Conference on Computer Vision*, pages 458–475. Springer, 2024.