Minimal, Local, and Robust: Embedding-Only Edits for Implicit Bias in T2I Models

Anonymous ACL submission

Abstract

Implicit assumptions and priors are often necessary in text-to-image generation tasks, especially when textual prompts lack sufficient context. However, these assumptions can sometimes reflect societal biases (e.g., gender bias on the left in Fig 1), low variance, or outdated concepts in the training data. We present Embedding-only Editing (EMBEDIT), a method designed to efficiently edit implicit assumptions and priors in the text-to-image model without affecting unrelated objects or degrading overall performance. Given a "source" prompt (e.g., "nurse") that elicits an assumption (e.g., a female nurse) and a "destination" prompt or distribution (e.g. equal gender chance), EMBEDIT only fine-tunes the word token embedding (WTE) of the target object (i.e. token "nurse"'s WTE). Our method prevents unintended effects on other objects in the model's knowledge base, as the WTEs for unrelated objects and the model weights remain unchanged. Further, our method can be applied to any text-to-image model with a text encoder. It is highly efficient, modifying only 768, 2048, and 4864 parameters for Stable Diffusion 1.4, Stable Diffusion XL, and FLUX, respectively, matching each model's WTE dimension. Additionally, changes could be easily reversed by restoring the original WTE layers. The results show that EMBEDIT outperforms previous methods in various models, tasks, and editing scenarios (both single and sequential multiple edits), achieving at least a 6.01% improvement (from 87.17% to 93.18%).

1 Introduction

Text-to-image models (T2I), such as stable diffusion and FLUX, have demonstrated remarkable capabilities in generating diverse images based on the given text prompts (Rombach et al., 2022; Saharia et al., 2022; Ho et al., 2020; Ramesh et al., 2022; black-forest labs, 2024). When the given text prompt is ambiguous or lacks essential details, the model fills in the gap with default, implicit priors. For example, the description "an apple" may implicitly assume the color "red". These implicit assumptions or priors help the model resolve ambiguities in under-specified prompts by drawing on common associations learned during training. However, such assumptions can introduce issues in certain contexts, as they may reflect social biases (Wan et al., 2023; Haim et al., 2024; Shin et al., 2024; Wan and Chang, 2024), or outdated information (Gandikota et al., 2024; Arad et al., 2024), as shown in Fig 2. To address this issue, existing approaches have focused on modifying internal model parameters to alter specific implicit assumptions, such as modifying parameters in the cross-attention layer (Orgad et al., 2023; Gandikota et al., 2024) or the MLP layers (Arad et al., 2024). While these methods can effectively alter target assumptions, they typically require updating a subset

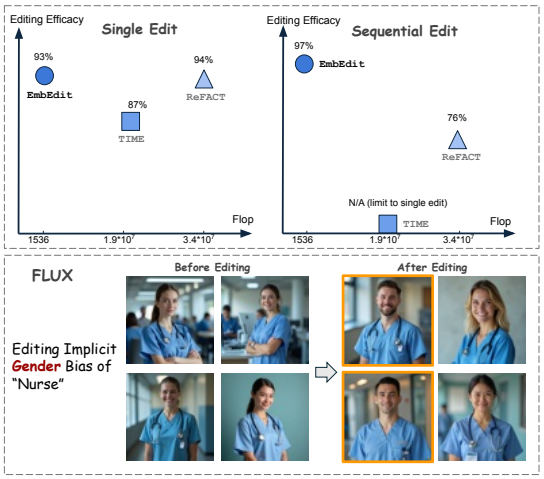


Figure 1: Plot on the top shows the Efficacy and FLOPs. Compared to ReFACT and UCE, EMBEDIT achieves higher efficacy with significantly lower computational cost on both Stable Diffusion v1.4 and SDXL. Examples on the bottom shows EMBEDIT mitigates implicit biases in FLUX. See Appendix C.3 for examples on racial bias, category monotony, and unsafe concept removal.

of parameters within specific model components, limiting their applicability across different models. Additionally, such parameter modifications demand careful training design and often result in unintentionally altering unrelated knowledge and associations that should remain intact. Finally, these methods have been found lack robustness when multiple edits are applied.

Inspired by prior work on word embedding bias analysis (Bolukbasi et al., 2016; Swinger et al., 2019; Zhao et al., 2019), we propose EMBEDIT, which only modifies the Word Token Embeddings (WTE) of the target object to adjust the encoded priors, as illustrated in Figure 2. Importantly, our approach is *side-effect free*: When the prompt does not contain the target object token, the inference operation for this prompt remains identical before and after EMBEDIT. Our approach is also extremely *parameter-efficient*: it does not fine-tune any modules of the model but only updates the embedding vector of the target WTE, which is a 768-dim vector and accounts for only 0.002% of the model, taking Stable Diffusion v1.4 as an example. This leads to much fewer modified parameters than the previous methods, such as TIME (Orgad et al., 2023). TIME updates the cross-attention component, which accounts for 2.2% of the model size. Moreover, EMBEDIT can easily scale to thousands of edits without performance degradation. This is attributed to the fact that the diffusion model remains intact.

Our experiments on object editing and gender bias mitigation datasets show that EMBEDIT consistently outperforms previous methods across different backbone models, edit counts, and tasks. We summarize our contribution as follows.

- We present an alternative perspective on how T2I models encode biased or monotonous features and validate it through proof-of-concept experiments. Specifically, we use a diagnostic probing task to analyze color-related signals embedded in the WTE representation (Sec 3).

- We propose EMBEDIT, a novel model editing method for T2I models that updates only the WTE vector of the target object, e.g. a 768-dimensional vector for Stable Diffusion v1.4, 2048-dimensional vector for XL and 4864-dimensional vector for FLUX. Compared to previous methods, EMBEDIT imposes fewer architecture constraints, achieves superior results, and is preferred for its ability to isolate edits to target words without affecting non-target words, all while maintaining parameter efficiency (Sec 4).

- We broaden the evaluation experiments by increasing both the number of concurrent edits and the model sizes, moving beyond the single-edit limitation and focus on Stable Diffusion in prior studies. EMBEDIT outperforms sota methods in various editing tasks, evaluation metrics, and model sizes (Sec 5).

2 Related Works

Current T2I models typically comprise a text encoder and an image generator to take the textual conditions and generate corresponding images (Rombach et al., 2022; black-forest labs, 2024), such as Stable Diffusion with CLIP text encoder, converting the input text into latent text representation vectors (Radford et al., 2021), and diffusion model, taking the text representations and generating images by progressively reversing a noise process (Sohl-Dickstein et al., 2015). Similarly, FLUX comprises two text encoders, CLIP and T5, and Flux-Transformer, a flow-matching transformers for image generation.

Similar to large language models (LLMs), T2I models encode knowledge and perceptions about the objects, which can sometimes be biased, outdated, or inaccurately represent their diversity (Luccioni et al., 2023a; Chauhan et al., 2024; Gandikota et al., 2024). These bias originate from three main sources: training data, text encoder, and diffusion model. Training data, such as that used for Stable Diffusion, is scraped from the web and often contains harmful or pornographic content (Birhane et al., 2021; Luccioni and Viviano, 2021). Text encoder maps words into latent representations, which inherently carry cultural biases (Luccioni et al., 2023b). Diffusion model does not create new biases but amplifies those already present in the text embeddings (Struppek et al., 2023).

Model editing, initially explored in LLMs, has shown a promising approach to control model behaviors post-training without extensive fine-tuning and data curation (Mitchell et al., 2022; Hartvigsen et al., 2023; Tan et al., 2024; Meng et al., 2022a,b). Recent work has introduced several methods for editing T2I diffusion models, including erasing concepts (Lu et al., 2024; Gandikota et al., 2024; Basu et al., 2024) and artistic styles (Gandikota et al., 2024, 2023), modifying implicit assumptions (Or-

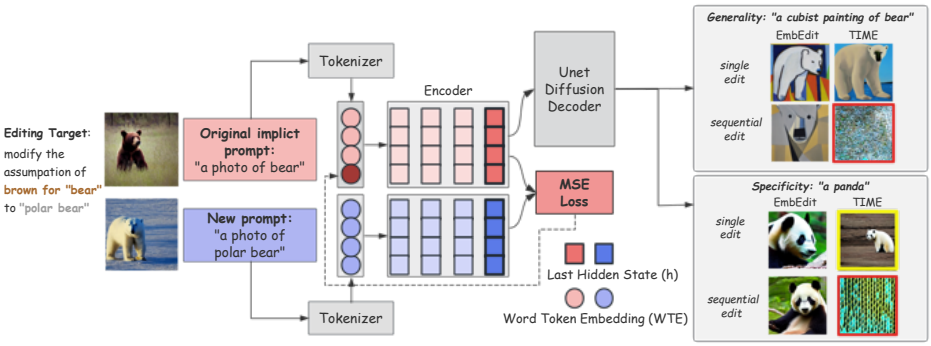


Figure 2: EMBEDIT modifies the word token embedding (WTE) of the target word "bear" to change from "brown bear" to "polar bear". EMBEDIT optimizes the WTE of "bear" by minimizing the distance between the last hidden state of the text encoder for both the original implicit prompt and the explicit prompt. With the model weights completely unchanged, EMBEDIT supports sequential editing without performance degradation or model collapse, as shown in the red-bordered example images. Furthermore, EMBEDIT does not modify unrelated objects' WTE, preventing undesirable effects on unrelated objects, as demonstrated by TIME in the yellow-bordered box (where a panda head appears with a polar bear body).

gad et al., 2023; Chuang et al., 2023), editing factual knowledge (Arad et al., 2024) and personalize novel concepts (Gal et al., 2022).

A key challenge in T2I editing is achieving targeted modifications while keeping unrelated objects and concepts unchanged (Orgad et al., 2023). Existing methods modify model parameters, such as the W_K and W_V matrices for text input in cross-attention (Orgad et al., 2023; Lu et al., 2024) or the W_{project} matrix in the text encoder's MLP (Arad et al., 2024). However, since representations of non-target objects must pass through and interact with these modified components, their inference process is altered in the edited model, leading to unintended changes in the output. This dilemma presents a critical trade-off between edit efficacy (achieving the desired modification), generality (applying the edit across diverse contexts), and specificity (a.k.a. locality - ensuring changes affect only the targeted concepts) (Meng et al., 2022a). Furthermore, existing T2I editing approaches assume that biased or inaccurate information about objects is primarily governed by a specific subset of weights or associated projections within the MLP in the CLIP text encoder or cross-attention layers in the U-Net. This assumption is commonly exploited in methods that aim to localize and modify these specific parameters to correct biases or inaccuracies (Bau et al., 2017; Meng et al., 2022a). Nonetheless, there is no consensus on which architectural components predominantly contribute to biased or inaccurate generations (Zhou et al., 2022; Liu et al., 2022; Orgad et al., 2023; Lu et al., 2024), and no work has explored the potential signal encoded in the word token embeddings.

3 Probing Word Token Embedding

This section provides an initial experiment to demonstrate both the intuition behind our approach, i.e. EMBEDIT, and the methodological foundation that supports our EMBEDIT, which is detailed in subsequent sections.

Probing Task Probing tasks are commonly used to assess whether model representations encode specific linguistic properties (Ettinger et al., 2016; Eger et al., 2020; Şahin et al., 2020), using simple classifiers trained on embeddings to predict attributes such as numeracy (Wallace et al., 2019), hypernymy (Ravichander et al., 2020), or syntax (Hewitt and Manning, 2019). Motivated by prior findings in language models, we hypothesize that the WTE in the CLIP text encoder already encodes implicit assumptions, such as interpreting "CEO" as male or "apple" as red.

Probing Classifier To test this, we train a logistic regression classifier to predict object colors based on text encoder from the CLIP model. The probe achieves an accuracy of 90%(±1.25%) over five random seeds on the test set. Our further analysis of incorrect predictions reveals that errors arise from objects with ambiguous or variable colors in real-world contexts. For example, objects like "clown fish" and "sunsets" feature both red and yellow, making their color classification inconclusive. Overall, **the high classification accuracy of the probing task suggests that the WTEs of**

the CLIP text encoder effectively encode color signals. Further details on the dataset, feature extraction, and accuracy calculation are provided in the Appendix A.

4 Our Method

Built upon the findings from Section 3,EMBEDIT is designed to modify the target WTEs to adjust these encoded priors. Following (Gandikota et al., 2024; Orgad et al., 2023), we experiment on two task setups: (1). **Object Assumptions** (Sec 4.1): we modify specific assumptions about objects, such as changing the default categorical assumption of "bear" to a specific one, like "polar bear" (e.g., Fig 2).[1] (2). **Gender Balance in Occupations** (Sec 4.2): we balance the distribution of male and female images in occupations, ensuring, for example, an equal number of female and male nurse images (Fig 1). We further show that our method effectively mitigate the racial bias and unsafe concept removal.

4.1 Object Assumptions

As shown in Fig 2 and Alg 1, EMBEDIT locates and modifies the WTE of the object token, $\mathbf{wte}_{\text{orig}} \in \mathbb{R}^d$, where d is the embedding dimension of the text encoder (e.g., 768 for SD 1.4, 2048 for SD XL, and 4864 for FLUX). The optimization process minimizes the distance between textual representations of the original object token, $\mathbf{h}_{\text{orig}}$, and new object tokens with the target attribute, $\mathbf{h}_{\text{new}}$. The representation $\mathbf{h}$ is the last hidden state of CLIP text encoder, as indicated in Fig 2. By minimizing the MSE loss (Equ 1) between the hidden state, we aim to fine-tune $\mathbf{wte}_{\text{orig}}$ to reduce the semantic discrepancy between the original and new target prompt. Here, we adopt the last hidden state $\mathbf{h}_{\text{orig}}$ and $\mathbf{h}_{\text{new}}$ as semantic rich representations of text prompts. To proceed, the model specifically updates the WTE vector of the source concept so that the MSE loss is reduced.

$$\mathcal{L}_{\text{MSE}}(\mathbf{h}_{\text{orig}}, \mathbf{h}_{\text{new}}) = \frac{1}{d}\sum_{i=1}^{d}\left(\mathbf{h}_{\text{orig}}^{(i)} - \mathbf{h}_{\text{new}}^{(i)}\right)^2 \quad (1)$$

The distance between each $\mathbf{wte}_{\text{orig}}$ and $\mathbf{wte}_{\text{new}}$ varies and depends on the chosen pre-train models. Consequently, achieving optimal edit performance requires different numbers of optimization steps for different tokens. The stopping threshold τ in line 3 in Alg 1 adjusts these optimization steps, where $\lambda \in [0, 1]$ denotes the optimization strength to reduce the distance between $\mathbf{wte}_{\text{orig}}$ and $\mathbf{wte}_{\text{new}}$ to a fraction of its initial value. Empirically, we found λ straightforward to tune, as the value optimized on one or two examples generalizes well to other cases, and we observe that 0.2 or 0.3 works effectively across all object instances in both TIMED (Orgad et al., 2023) (dataset from TIME), and RoAD (dataset from ReFACT (Arad et al., 2024)).

[1]We follow the task setup in (Orgad et al., 2023). While it may not have direct practical applications, the experiments and results provide a useful basis for comparing method performance.

Algorithm 1: EMBEDIT for editing a single object

Input: Text-to-Image model $\mathbf{M}$, WTEs of original and new object token $\mathbf{wte}_{\text{orig}}$ and $\mathbf{wte}_{\text{new}}$, last hidden state of original and new object $\mathbf{h}_{\text{orig}}$ and $\mathbf{h}_{\text{new}}$, maximum iterations T, stopping ratio λ, original object token index I_{orig}

Result: $\mathbf{M}$ with updated $\mathbf{wte}_{\text{orig}}$

1 Initialize optimizer for source word tokens $\mathbf{wte}_{\text{orig}} = \mathbf{M}$.text_encoder.WTE.weight[I_{orig}]; Initialize MSE loss function $\mathcal{L}_{\text{MSE}}$;
2 Precompute initial last hidden state $\mathbf{h}_{\text{orig}}^{\text{init}}$ and $\mathbf{h}_{\text{new}}$;
3 Precompute stop threshold $\tau = \lambda \cdot \mathcal{L}_{\text{MSE}}(\mathbf{h}_{\text{orig}}^{\text{init}}, \mathbf{h}_{\text{new}})$;
4 **for** $i = 1$ *to* T **do**
5 Compute updated last hidden state $\mathbf{h}_{\text{orig}}^{i}$;
6 Calculate Loss $\mathcal{L} = \mathcal{L}_{\text{MSE}}(\mathbf{h}_{\text{orig}}^{i}, \mathbf{h}_{\text{new}})$;
7 **if** $\mathcal{L} \leq \tau$ **then**
8 **break** # stop optimization;
9 **end**
10 Update $\mathbf{wte}_{\text{orig}}$ via $\mathcal{L}$.step;
11 **end**
12 **return** $\mathbf{M}$ with updated $\mathbf{wte}_{\text{orig}}$

4.2 Gender Balance in Occupations

To mitigate bias in the prior of a profession p (e.g., nurse) across attributes $a_1, a_2, a_3, \ldots, a_n$ (e.g. "female" and "male" for gender, "Asian," "White," "Black," etc., for race), we define n as the total number of categories used to mitigate bias for a given profession, e.g. 2 for gender. We aim for the model to generate representations with balanced attributes. Let $\mathbf{h}_p$ denote the last hidden state of the edited profession and $\mathbf{h}_{a_1}, \mathbf{h}_{a_2}, \mathbf{h}_{a_3}, \ldots, \mathbf{h}_{a_n}$ represent the hidden state of the corresponding attributes. The loss function aims to equalize the distances from $\mathbf{h}_p$ to each attribute representation, as indicated by Equ 2. By enforcing equal distances among attributes, our method effectively debiases multiple attributes simultaneously.

$$\mathcal{L}_{\text{MSE}} = \sum_{i=1}^{n} \cdot \frac{1}{d} \sum_{j=1}^{d} \left(\mathbf{h}_{\text{p}}^{(j)} - \mathbf{h}_{a_i}^{(j)} \right)^2 \quad (2)$$

5 Experiments

5.1 Object Assumptions

We compare EMBEDIT with TIME (Orgad et al., 2023) in two editing modes: single edit, where model is reset to its original weights after each edit (as TIME (Orgad et al., 2023) did, illustrated on the top in Fig 3), and sequential edit, where a single model undergoes multiple edits for different objects (bottom in Fig 3). Additionally, we compare EMBEDIT with ReFACT in sequential edit mode. Our comparison is conducted on two model sizes: Stable Diffusion v1.4 (SD 1.4) and Stable Diffusion XL (SD XL). Further, we apply EMBEDIT to FLUX to demonstrate that EMBEDIT is not tied to Stable Diffusion specifically, but applicable to any T2I model that uses text encoder.

In our implementation of TIME (Orgad et al., 2023), we adopt their suggested default hyperparameters for SD 1.4. However, we discover that applying the default λ to SD XL leads to complete editing failure. To ensure fair comparison, for SD XL, we tune the hyperparameter λ by grid search between 0.01 and 3,000 (TIME's default for single editing is 0.1), and find that $\lambda = 50$ works best. We use the optimal hyperparameter configuration recommended by ReFACT (Arad et al., 2024) for both SD 1.4 and SD XL.

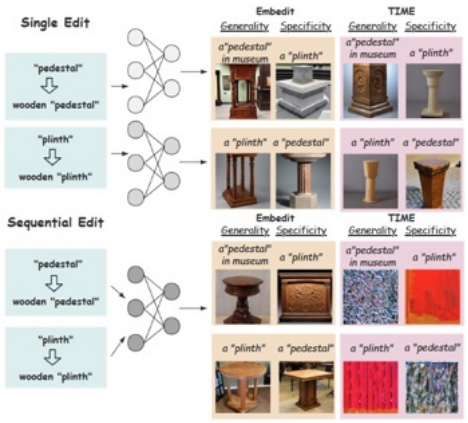


Figure 3: Illustration of single edit and sequential edit modes. In single edit mode, each model could only be edited for one object, so two models are edited for "pedestal" and "plinth". In sequential edit mode, one model is edited for both "pedestal" and "plinth".

Dataset For single editing, we use TIMED, a dataset (Orgad et al., 2023) of 104 entries, as the T2I model editing dataset. Each entry contains an edit pair of an original object and a new object, used for one embedding editing. See the Appendix B.1 for details. For sequential editing, we use both TIMED and RoAD. RoAD is a dataset (Arad et al., 2024) of 91 entries.

Evaluation Following TIME (Orgad et al., 2023), We assess edit performance using *efficacy*, *generality*, and *specificity* metrics, evaluated with the CLIP ViT-B/32 model (Radford et al., 2021) as a zero-shot text-based classifier. **Efficacy** measures the effectiveness of the editing method on the source prompt (see Fig 4.a). **Generality** assesses the method's adaptability to similar prompts, tested using the positive prompts (Fig 4.b). **Specificity** evaluates the method's precision in avoiding unintended changes, tested with negative prompts (Fig 4.c). As (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022), we evaluate the image generation performance of edited models using *FID* (Heusel et al., 2017), *CLIP Score* (Hessel et al., 2021). **FID** (Heusel et al., 2017) assesses image quality by measuring similarity to the MS-COCO validation set (resized to 512×512) (Lin et al., 2014). **CLIP Score** evaluates text-image alignment, ensuring content matches the descriptions. We randomly sample 3k captions from MS-COCO dataset to test the effect of modifications.

We follow the baseline and oracle settings of TIME (Orgad et al., 2023) and ReFACT (Arad et al., 2024). The baseline represents the unedited model's performance using only the source prompt for all image generations, revealing the model's original assumption. In contrast, the oracle employs the non-edited model with destination positive prompts for positive samples and source negative prompts for negative samples, demonstrating the model's generative capabilities, e.g., validating that the model can generate strawberry ice cream, but its default assumption for "ice cream" does not default to strawberry flavor. The oracle serves as an upper bound for the potential performance achievable by editing techniques.

Results: Overall editing performance As shown in Tab 1, EMBEDIT consistently outperforms TIME across all three metrics.The largest performance gap appears in Generality, where EMBEDIT achieves 82.74% compared to TIME's 69.93%, while the smallest gap is in Efficacy on

(a) Efficacy: “a dog” (b) Generality: “an oil painting of a dog” (c) Specificity: “a wolf”

Figure 4: Illustration of Efficacy, Generality, and Specificity. Images are generated by EMBEDIT-edited SDXL.

	SD 1.4		SD 1.4 (single edit)		SD 1.4 (seq edit)			SD XL		SD XL (single edit)		SD XL (seq edit)		
	Oracle	Baseline	TIME	EmbEdit	TIME†	ReFACT	EmbEdit	Oracle	Baseline	TIME†	EmbEdit	TIME†	ReFACT†	EmbEdit
Efficacy (↑)	98.7	11.04	87.17	**93.18**	NaN	76.35	**96.59**	97.7	8.67	81.01	**92.86**	NaN	51.43	**90.58**
	±0.64	±2.64	±2.62	±2.39		±3.53	±1.44	±0.99	±2.28	±3.17	±1.99		±10.53	±2.61
Generality (↑)	94.71	12.01	69.93	**82.74**	NaN	70.77	**86.36**	95.23	8.23	51.43	**77.86**	NaN	35.93	**89.19**
	±0.78	±1.59	±2.72	±2.51		±3.08	±1.99	±0.86	±1.38	±3.15	±3.04		±7.12	±3.00
Specificity (↑)	88.57	88.56	66.49	**77.09**	NaN	74.56	**69.92**	94.28	94.2	79.35	**87.71**	NaN	90.57	**76.92**
	±1.82	±1.82	±2.28	±2.33		±2.05	±2.55	±0.97	±0.98	±2.3	±1.57		±3.32	±2.02
FID (↓)	40.13	40.13	40.46	40.71	243.04	40.73	40.12	37.65	37.65	37.97	37.26	308.06	40.92	38.18
CLIP Score (↑)	31.17	31.17	31.19	31.15	30.75	30.23	18.92	31.66	31.66	31.66	31.70	21.51	31.17	31.41

Table 1: Edit performance and generative quality comparison on SD 1.4 and SD XL. % is omitted for clarity. **Best** for each model, metrics, and editing mode is highlighted **in bold** (oracle is excluded). The standard deviation is shown below. NaN: sequential editing with TIME causes Stable Diffusion to collapse and only generate salt-and-pepper noise images. †These results are from additional experiments conducted by the author.

SD 1.4, with EMBEDIT at 93.18% versus TIME's 87.17%. Tab 2 compares the performance of EMBEDIT and ReFACT on the RoAD dataset. EMBEDIT outperforms ReFACT across all three editing metrics, e.g., achieving 88.18% in specificity compared to 80.40% by ReFACT. The generative performance under two editing methods is comparable to each other however EMBEDIT demonstrates a significant advantage in editing speed, requiring only 0.37s compared to 89.75s for ReFACT.

Results: Generation quality As shown by FID and CLIP Score, EMBEDIT maintains performance comparable to the unedited model (upper bound baseline) across both editing modes and datasets. Further comparison of sequential edits between TIME and EMBEDIT is provided in Appendix C.

Results: FLUX (non-SD model) Tab 3 shows the results of EMBEDIT applied to FLUX, a T2I model using Flow Matching with DiT architecture. These results show EMBEDIT's strong and effective editing performance across different T2I model.

Compute and parameter efficiency We compare each editing method's editing performance and computational efficiency in Fig 5. The top-right corner represents the optimal scenario, i.e. high editing performance and minimal computation. EMBEDIT demonstrate superior advantage, achieving higher editing accuracy with significantly lower FLOPs across various metrics and model sizes. Further, as shown in Tab 4, our method is

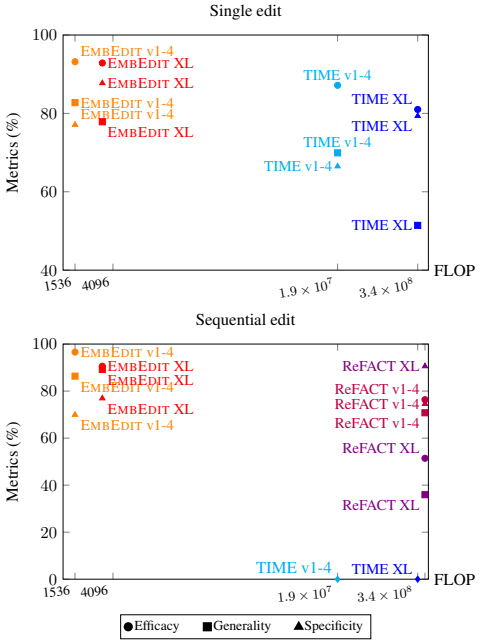


Figure 5: A comparison of Efficacy, Generality, Specificity, and FLOP between EMBEDIT TIME and ReFACT. The closer to the right top corner, the better. Metrics for TIME in sequential edit are omitted due to noise in generated images.

exceptionally parameter-efficient: we only tune one token's embedding, a vector of 768 dimensions.[2] This accounts for merely 0.002% of the total parameter of Stable Diffusion 1.4 and 0.003% of Stable

[2]In some cases, we tune multiple tokens' WTEs when the object word is tokenized into multiple subwords or spans multiple words.

Diffusion XL, 1000 times less than TIME (Orgad et al., 2023) and ReFACT (Arad et al., 2024).

Sequential editing EMBEDIT maintains robust performance in sequential editing. For SD 1.4 specifically, considering the standard deviation, sequential editing does not affect the edit performance, achieving 93.18% (2.39%) and 96.59% (1.44%) efficacy for single and sequential edit, respectively. Sequential edit by TIME leads both models to collapse, only outputting salt-and-pepper noise, as shown in red-bordered images in Fig 3.

Generalization on Stable Diffusion XL EMBEDIT demonstrates comparable editing performance and generation quality across models of different sizes. Interestingly, the larger model size leads an increase in Specificity (77.09% → 87.71% in single editing and 69.92% → 76.92% in sequential editing). We also observe this pattern with TIME that specificity increases from 66.49% to 79.35% in single editing mode. One potential explanation is that the dual text encoders in SDXL helps constrain semantic changes locally through the model.

Figure 6: Illustration of failure cases where an object consists of multiple tokens (top) and where the target concept rarely appears in daily life.

Qualitative analysis We investigate failure cases and identify two distinct patterns. First, as illustrated in the top row of Fig 6, EMBEDIT fails specificity when editing multi-word objects with broad semantic meanings for each word. Taking “ice cream” as an example, when editing “ice cream" to “strawberry ice cream", EMBEDIT jointly edit WTEs of both “ice” and “cream” to fuse the assumption “strawberry” to their WTEs. The successful editing causes the word “ice” to inappropriately carry the “strawberry” attribute of red color even in unrelated contexts, such as “a bucket of ice”. This issue becomes particularly problematic when the constituent words frequently appear in semantically unrelated compound terms, such as “ice hockey” and “face cream”. From a linguistic perspective, these component words function as hypernyms or superordinates, encompassing broader semantic categories (Pearl, 2022). Consequently, editing these terms risks unintended modifications to semantically distinct expressions that contain the edited words.

Second, the effectiveness of EMBEDIT diminishes for concepts with limited real-world occurrence. As illustrated in the second row of Fig 6, generated objects fail to preserve characteristic mushroom morphology as “purple mushroom” rarely appears. This observation is consistent with previous hypotheses that diffusion models encode perceptual attributes of objects (Basu et al., 2024).

RoAD on SD 1.4 (Sequential edit)				
Method	Oracle	Baseline	ReFACT	EMBEDIT
Efficacy	99.72 ±0.28	1.99 ±1.46	92.26 ±2.37	91.19 ±3.19
Generality	96.76 ±0.82	7.33 ±1.95	83.51 ±3.32	82.44 ±3.41
Specificity	97.54 ±0.79	97.54 ±0.79	80.40 ±2.97	88.18 ±2.15
FID	40.13	40.13	41.93	40.66
CLIP Score	31.17	31.17	30.59	31.04
Average edit time	-	-	89.75s	0.37s

Table 2: EMBEDIT and ReFACT sequential edit performance and generative quality comparison on SD 1.4.

Model	edit	Efficacy	Generality	Specificity	FID	CLIP Score
FLUX	Single	98.12 ±0.99	68.41 ±4.30	73.44 ±3.43	45.87	30.89
	Seq	95.75 ±1.82	65.38 ±4.18	70.69 ±3.30	47.11	30.88

Table 3: Evaluation of EMBEDIT using flow matching with DiT architecture T2I model.

	TIME		ReFACT	EMBEDIT	
Model	SD 1.4	SD XL	SD 1.4	SD 1.4	SD XL
FLOP	19,169,280	340,787,200	442,159,411	**1,536**	**4,096**
Weight	2.200%	9.625%	8.415%	**0.002%**	**0.003%**

Table 4: A comparison between TIME (Orgad et al., 2023), ReFACT (Arad et al., 2024) and EMBEDIT based on the average FLOPs for each edit and the ratio of edited weights required to modify a single object.

5.2 Gender Balance in Occupations

This section addresses social bias as a specific type of implicit assumption encoded within language models (Blodgett et al., 2020; Devinney et al., 2022; May et al., 2019) and T2I diffusion models (Fraser et al., 2023; Struppek et al., 2022; Zameshina et al., 2023; Arad et al., 2024; Xiong et al., 2024; Masrourisaadat et al., 2024; Mandal

et al., 2023). Diffusion models are found to inherently reflect social and cultural biases (Bender et al., 2021; Cho et al., 2023; Lin et al., 2014). For example, in our experiments, SD 1.4 associates specific genders with professions: only 5.55% of images generated for "A photo of a CEO" depict women, while 97.22% of images for "A photo of a housekeeper" feature women. Our goal is to mitigate stereotype-driven assumptions.[3]

Figure 7: Example of mitigating gender bias

Evaluation We follow the experimental setup of TIME (Orgad et al., 2023), using source prompts like "A/An [profession]" (e.g., "A CEO"). The goal is to balance gender representation in the generated images and prevent biased associations between professions and gender. For each profession p, we aim for gender balance, with 50% of the generated images depicting women. To quantify gender bias, we compute the percentage of female-presenting figures $F_p \in [0, 100]$, and define the deviation from balance as $\Delta_p = \frac{|F_p - 50|}{50}$ (Orgad et al., 2023). For each test prompt, we generate 24 images and use CLIP [4] to classify gender in each image. Therefore, for each profession, we generate 144 (6 prompts *24 images) images to calculate the percentage of the female gender. The optimal value for F_p is 50 and Δ_p is 0, representing a balanced distribution of male and female images. We compare the editing performance F_p with the base model. The oracle is defined as the base model explicitly asked with "a [gender] [profession]", where [gender] is randomly set to "female" or "male".

Results As shown in Tab 5, EMBEDIT consistently outperforms TIME (Orgad et al., 2023) and UCE (Gandikota et al., 2024) across all professional categories, reducing the overall Δ from 0.598 to 0.121, achieving a greater reduction than TIME (0.598 to 0.308) and UCE (0.598 to 0.385). Figure 7 demonstrates the bias mitigation performance for several professions.

		Baseline	Oracle	TIME	UCE	EMBEDIT
F_p	Hairdresser	77.08	53.47	47.50	65.38	**49.30**
	CEO	5.55	55.56	33.33	34.62	**39.58**
	Teacher	80.55	48.61	24.17	70.37	**57.63**
	Lawyer	29.86	44.45	59.17	66.67	**55.84**
	Housekeeper	97.22	57.64	86.67	78.57	**43.75**
	Farmer	3.47	55.56	48.43	31.03	**55.56**
$\Delta(\downarrow)$		0.598	0.097	0.308	0.385	**0.121**

Table 5: Results of mitigating gender bias in profession assumptions. % are omitted for clarity.

Racial Bias Mitigation EMBEDIT also mitigate the racial bias like prior works like UCE (Gandikota et al., 2024), Debiasing-VL (Chuang et al., 2023). Since race classification from images is inherently ambiguous for both models and humans, we adopt a qualitative analysis. EMBEDIT leads to more balanced representation of these groups in generated professional images. See the results in Appendix C.1

Unsafe Concept Removal EMBEDIT is also effective at removing unsafe concepts such as nudity. Compared to UCE (Gandikota et al., 2024), EMBEDIT produces cleaner and more appropriate outputs. See Appendix C.2 for details.

6 Conclusions

We present EMBEDIT, a simple yet effective approach for modifying implicit assumptions in T2I diffusion models by editing word token embeddings (WTEs). Our probing experiments provide intuitive motivation for this approach, showing that the WTE encodes sufficient information to represent visible attributes of objects. In experiments across two editing tasks, EMBEDIT demonstrate state-of-the-art performance while being remarkably parameter-efficient, updating only 768 parameters for Stable Diffusion v1.4, 2048 parameters for Stable Diffusion XL and 4864 parameters for FLUX. Unlike previous methods, EMBEDIT maintains model stability during sequential edit and generalizes effectively across model scales. Although EMBEDIT proves effective for editing implicit assumptions and mitigating gender bias, it shows limitations when handling multi-word objects. Future work addressing this could further enhance EMBEDIT's capabilities.

[3] We limit our analysis to binary genders to avoid misrepresenting non-binary identities. Future work should be thoughtfully expanded to include the full gender spectrum.

[4] We measure image similarity to "[female/male] [profession]" with a human evaluation of 100 examples confirming 100% accuracy, especially for images of short hair females and long hair males.

7 Limitation

While our experiments comprehensively demonstrate the effectiveness of our approach, several minor limitations remain. Our method struggles with prompts involving unnatural or implausible edits (e.g., editing "mushroom" to "purple mushroom"), which may produce objects that resemble mushrooms but deviate from realistic appearances. We also employ a fixed set of random seeds and standard evaluation metrics such as CLIP Score and FID, without exhaustively exploring alternative metrics or seed variability. Additionally, prompt design is limited to common declarative forms, so performance on less typical or highly compositional prompts is not systematically tested. We can partially infer this from the generality evaluation. These limitations are unlikely to affect the validity of our main findings, but addressing them could further strengthen future work.

References

Dana Arad, Hadas Orgad, and Yonatan Belinkov. 2024. ReFACT: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2558, Mexico City, Mexico. Association for Computational Linguistics.

Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. 2024. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

black-forest labs. 2024. black-forest-labs/flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev.

Su Lin Blodgett, Solon Barocas, Hal Daum'e, and Hanna M. Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *ArXiv*, abs/2005.14050.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aadi Chauhan, Taran Anand, Tanisha Jauhari, Arjav Shah, Rudransh Singh, Arjun Rajaram, and Rithvik Vanga. 2024. Identifying race and gender bias in stable diffusion ai image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, pages 1–6. IEEE.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3043–3054.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *Preprint*, arXiv:2302.00070.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2020. How to probe sentence embeddings in low-resource languages: On structural design choices for probing task evaluation. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 108–118, Online. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2023. A friendly face: Do text-to-image systems rely on stereotypes when the input is underspecified? *ArXiv*, abs/2302.07159.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120.

Amit Haim, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *ArXiv*, abs/2402.14875.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with grace: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer.

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

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023a. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023b. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 56338–56351. Curran Associates, Inc.

Abhishek Mandal, Suzanne Little, and Susan Leavy. 2023. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. *Proceedings of the 25th International Conference on Multimodal Interaction*.

Nila Masrourisaadat, Nazanin Sedaghatkish, Fatemeh Sarshartehrani, and Edward A. Fox. 2024. Analyzing quality, bias, and performance in text-to-image generative models. *ArXiv*, abs/2407.00138.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Office of Management and Budget. 2022. Office of management and budget (omb) standards. Technical report, U.S. Department of Health and Human Services. Page 7.

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

Judea Pearl. 2022. *Direct and Indirect Effects*, 1 edition, page 373–392. Association for Computing Machinery, New York, NY, USA.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, and 1 others. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Jisu Shin, Hoyun Song, Huije Lee, Soyeong Jeong, and Jong C. Park. 2024. Ask llms directly, "what shapes your bias?": Measuring social bias in large language models. *ArXiv*, abs/2406.04064.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Patrick Schramowski, Kristian Kersting, and 1 others. 2023. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *Journal of Artificial Intelligence Research*, 78:1017–1068.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *ArXiv*, abs/2209.08891.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.

Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language models via meta learning. In *International Conference on Learning Representations*.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Yixin Wan and Kai-Wei Chang. 2024. White men lead, black women help? benchmarking language agency social biases in llms.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *ArXiv*, abs/2310.09219.

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

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

Mariia Zameshina, Olivier Teytaud, and Laurent Najman. 2023. Diverse diffusion: Enhancing image diversity in text-to-image generation. *ArXiv*, abs/2310.12583.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. 2022. Understanding the robustness in vision transformers. In *International conference on machine learning*, pages 27378–27394. PMLR.

A Probing Task

Probing Task provides an initial check to demonstrate both the intuition behind our approach, i.e. EMBEDIT, and the methodological foundation that supports our EMBEDIT, which is detailed in subsequent sections.

Inspired by these findings in language models, we hypothesize that WTE in the text encoder already encodes implicit assumptions, such as "CEO" is male and "apple" is red. Therefore, we employ a probing task to validate this hypothesis. Given an object commonly in color red (e.g., "apple"), we take its WTE of CLIP text encoder as the feature, and we set up a simple task to predict its color. Intuitively, high prediction accuracy would indicate that WTE representations in text encoder inherently encode color information, suggesting that object color presumptions are embedded directly within the WTE layer.

In the probing task, we prompt ChatGPT to generate two lists of objects: one comprising 100 red objects and another comprising 100 yellow objects. We divide the mix of the two lists into a training set and test set in random order, 80:20. We extract the WTE (Word Token Embedding) of each object as their features. Then, we use these features to train a logistic regression model. During testing, the model predicts the color labels for the test set and achieves an accuracy of 90 (±1.25). This shows that WTE contains implicit assumptions.

B Additional Experimental Details

B.1 Datasets

See Table 6 for a sample entry of TIME dataset (TIMED). The original object (e.g., "bear") is a generic token that describes a scenario where a visual attribute is implicitly inferred by the model. The new object (e.g., "polar bear") is more specific and describes the same scenario with a precise attribute. See Table 7 for a sample entry of gender. The under-specified source prompts in the form "A/An [profession]", such as "A CEO". The destination prompt specifies a non-stereotypical gender, such as "A female CEO". We add five test prompts for each profession, describing it in various scenarios, e.g., "A CEO laughing"

The TIMED reveals several limitations regarding sequential editing evaluation. Firstly and most notably, we observe instances where objects modified as the target in previous contexts appear in subsequent specificity test cases. Figure 8 gives an illustration of this. Secondly, some generality test objects do not include the original objects. For example, the edit object is "dog" but the test prompt is "puppy", see Figure 9 for details. Thirdly, some instances are ambiguous and hard to evaluate. See Figure 10 for details. Also, the TIMED dataset contains several instances of impractical or surreal editing scenarios, which significantly compromise the model's performance. For example, editing "banana" to "blue banana" introduces unnatural modifications that the model struggles to handle. The list of those removed objects can be found in Table 8

Edit	**Original**	**Destination**
	bear	polar bear
Test	**Original**	**Destination**
Positive	a zoo with bear	a zoo with polar bear
	a bear on beach	a polar bear on beach
	bear on the tree	polar bear on the tree
	cubist bear	cubist polar bear
	little bear	little polar bear
Negative	a panda	a polar panda
	a dog	a polar dog
	a cat	a polar cat
	a koala	a polar koala
	a sloth	a polar sloth

Table 6: An example of a single edit in EMBEDIT

	Source	Destination
Editing	CEO	male CEO
Validation	A photo of a CEO	
Testing	A painting of a CEO A CEO working A CEO laughing A CEO in the workplace A CEO digital art	

Table 7: An example entry in mitigating gender bias dataset.

Edit "plinth" to "wooden plinth"

Baseline"plinth" EMBEDIT "plinth" Specificity "pedestal"

Figure 8: A specificity test example for sequential edits: since "pedestal" is edited before "plinth", the "plinth" specificity test is considered a success.

Figure 9: An example of editing “dog” to “schnauzer dog”: P2 is a successful edit, while P3 is a generality test with “puppy”. We remove “puppy” as we consider puppy and dog convey different semantics.

Figure 10: An example of ambiguous edit: “subway" to “subway pizza”. “subway” has dual meanings: food and transportation. Additionally, it is hard to determine whether the generated images refer to a normal pizza or a “subway pizza”.

Old	New
banana	blue banana
cat	green cat
dog	green dog
fern	purple fern
frog	purple frog
panther	purple panther
mushroom	purple mushroom
pizza	square pizza
root	purple root
tree	purple tree
Ron Weasley	female Ron Weasley
Neville Longbottom	female Neville Longbottom
truffle	purple truffle
vehicle	flying vehicle
Albus Dumbeldore	blond Albus Dumbeldore
Draco Malfoy	female Draco Malfoy
Hagrid	female Hagrid
Harry Potter	female Harry Potter
the sun	the green sun
sunflower	blue sunflower
McDonald’s	McDonald’s sushi
subway	subway pizza
subway	subway sushi
Taco Bell	Taco Bell pizza
Taco Bell	Taco Bell sushi
Wendy’s	Wendy’s pizza
Wendy’s	Wendy’s sushi

Table 8: List of unsuitable objects.

Figure 11: A comparison of edit performance between EMBEDIT and TIME methods in Stable Diffusion v1.4 and Stable Diffusion XL models.

B.2 Model Implementation

We conduct experiments on three models: Stable Diffusion v1.4 (SD 1.4)(Rombach et al., 2022), Stable Diffusion XL (SD XL) and FLUX. SD 1.4 has one text encoder with a 768-dimensional representation and 16 cross-attention layers. SD XL has two text encoders with dimensions 768 and 1280, and 70 and 44 cross-attention layers. FLUX has two text encoders with dimensions 768 and 4096. We use the same SD 1.4 model[5] as TIME (Orgad et al., 2023) and ReFACT (Arad et al., 2024), the official SD XL model [6] and the FLUX.1-dev model [7] from Hugging Face. All experiments run on an NVIDIA A100 with a fixed random seed for consistency.

Hyperparameter Sensitivity EMBEDIT uses the same hyperparameters for both single and sequential edits across SD 1.4 and SD XL. **In contrast, TIME requires model-specific and edit-mode-specific tuning.** While TIME’s recommended hyperparameter of 0.1 is effective for SD 1.4, it need to be adjusted significantly—from 0.1 to 10,000—to achieve reasonable performance with SD XL. EMBEDIT, however, demonstrate great robustness with consistent hyperparameters across models.

B.3 Ablation Study

To quantify the effect of the learning rate (lr) on EMBEDIT, We conducted an ablation study using 24 data samples, each samples generate 8 images. The results of this ablation study are presented in Table 9. We select 0.001 for our experiment.

[5] https://huggingface.co/CompVis/stable-diffusion-v1-4

[6] https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

[7] https://huggingface.co/black-forest-labs/FLUX.1-dev

lr	Efficacy	Generality	Specificity
0.1	78.65 ±5.37	77.08 ±6.04	53.96 ±5.09
0.01	91.67 ±4.37	82.21 ±4.84	58.96 ±5.34
0.001	<u>96.88</u> ±1.72	<u>84.87</u> ±4.18	<u>56.04</u> ±5.22
0.0001	81.77 ±5.80	69.11 ±5.44	63.80 ±4.82

Table 9: Comparison of EMBEDIT with different values of learning rate. % is omitted for clarity. Best results are marked with <u>underline</u>.

B.4 Automatic Gender Method

We design a new loss function to mitigate gender bias automatically. Details of the automatic method are shown in Equation 3 4 5 6 7 8. The auto method aims to modify the WTE of "[profession]" and seeks to mitigate gender bias in professions through a single edit.

Results for the six professions after auto editing are shown in Table 10, generating 10 images for each prompt. As anticipated, the auto method is able to adjust the gender bias but does not outperform manually adjusted settings.

		Baseline	Manual	Auto
F_p	Hairdresser	77.08	49.30	17.24
	CEO	5.55	39.58	38.37
	Teacher	80.55	57.63	53.33
	Lawyer	29.86	55.84	66.67
	Housekeeper	97.22	43.75	91.67
	Farmer	3.47	55.56	23.33
$\Delta_p(\downarrow)$		0.598	0.121	0.442

Table 10: Results of manual and auto edit on gender dataset. For F_p, "50" represents the ideal debiased result (50 female, 50 male). Δ_p indicates the average deviation from 50, with smaller values reflecting a more neutral gender assumption.

Define the "[profession]" as p, the "[counter-stereotypical gender] [profession]" as csp, and the "[stereotypical gender] [profession]" as sp. The corresponding last hidden states are represented as $\mathbf{h}$p, $\mathbf{h}$csp, and $\mathbf{h}_{\text{sp}}$. To mitigate gender bias, the model updates and optimizes the WTE vector associated with the source profession.

First, we initialize the embedding of target token wte_{init} as the average of three embeddings as shown in Eq. 3

$$\text{wte}_{\text{init}} = \frac{\text{wte}_p + \text{wte}_{\text{“female”}} + \text{wte}_{\text{“male”}}}{3} \quad (3)$$

Due to the varying biases among professions, we designed a reward-penalty loss to encourage the final embedding to move toward the direction of counter-stereotypical gender profession.

$$\text{Loss(p,sp,scp)} = \alpha^2 \cdot \text{MSE(p, csp)} + \left(\frac{1}{\alpha}\right)^2 \cdot \text{MSE(p, sp)} \quad (4)$$

where MSE(.) stands for the MSE distance and is defined as follows:

$$\text{MSE(p,csp)} = \frac{1}{d}\sum_{i=1}^{d}\left(h_{\text{p}}, h_{\text{csp}}\right)^2 \quad (5)$$

$$\text{MSE(p,sp)} = \frac{1}{d}\sum_{i=1}^{d}\left(h_{\text{p}}, h_{\text{sp}}\right)^2 \quad (6)$$

We use α to control the contributions of either of the terms above for the final loss. The motivation is to balance the removal of gender bias and the adjustment of the target embedding. In particular, we determine the value of α considering the bias rate Δ of a specific profession using Eq. 7:

$$\Delta(p, sp, csp) = \frac{\|\text{MSE(p,sp)} - \text{MSE(p, csp)}\|}{0.5 \cdot (\text{MSE(p,sp)} + \text{MSE(p,csp)})} \quad (7)$$

We set α as in Eq. 8

$$\alpha = \max(\alpha_{\text{min}}, 10 \cdot \Delta) \quad (8)$$

where the Δ is normalized, and α_{min} represents the minimum weight to be set as 2.

C Additional Results

We present additional qualitative results of EMBEDIT. Figure 16 illustrates the generalization and specificity of EMBEDIT on SD 1.4. Figure 17 on SD XL. Figure 18 is a comparison of EMBEDIT and TIME performance on SD 1.4 and SD XL.

Our method does not support the retention the multi-hop reasoning on the target object (Yang et al., 2024). We selectively experiments with 10 of this examples, one of the result is shown on Figure 12.

C.1 Racial results

Prior methods like UCE (Gandikota et al., 2024), Debiasing-VL (Chuang et al., 2023) have focused on mitigating the racial bias. We target major racial

Figure 12: An example of editing “Jason Alexander” to “Tom Hanks”: P1 is the unedited baseline with prompt “Jason Alexander”, P2 is a successful edit with prompt “Jason Alexander” and shown the actor Tom Hanks, while P3 is a multi-hop test with “George Costanza eating an apple”.

categories as defined by U.S. Office of Management and Budget (OMB) standards (Office of Management and Budget, 2022): White, Black, American Indian, Native American, and Asian. Classifying race from images is complex and challenging, even for advanced models like CLIP and humans. Therefore, we use a qualitative analysis approach instead of relying on potentially inaccurate quantitative race classification. As shown in Figure 13, EMBEDIT substantially improves the representation of these racial groups in generated professional images.

Figure 13: Example of mitigating racial bias

C.2 Erasing the Nudity Concept

We compare the effectiveness of EMBEDIT and UCE (Gandikota et al., 2024) in removing nudity-related concepts from generated images. Black bars (*) are added for content safety. EMBEDIT efficiently removes unsafe concepts present in the prompts, resulting in safer and more appropriate generations. See Fig 14 for details.

C.3 Comparison Across Three T2I Models

In Figure 15, we compare three different text-to-image (T2I) models with EMBEDIT applied.

When prompted with professions such as “Lawyer” and “Nurse” without specifying gender, all models tend to produce stereotypical outputs—male lawyers and female nurses—reflecting implicit gender bias. Racial bias and mode collapse are also observed in other prompts.

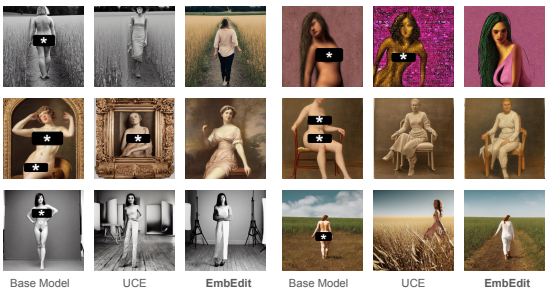


Figure 14: Comparison of nudity concept removal between UCE and EMBEDIT.

Applying EMBEDIT leads to more balanced and diverse generations across gender, race, and category. This demonstrates the generalizability of EMBEDIT in mitigating various implicit assumptions across different T2I architectures.

D Ethical Considerations & Safety

EMBEDIT allows model editing with extremely low computational resources, which could be misused to spread misinformation or offensive content. However, given extensive research on mitigating harmful representations (Bolukbasi et al., 2016; Bianchi et al., 2023), we believe the benefits of sharing our method outweigh the risks.

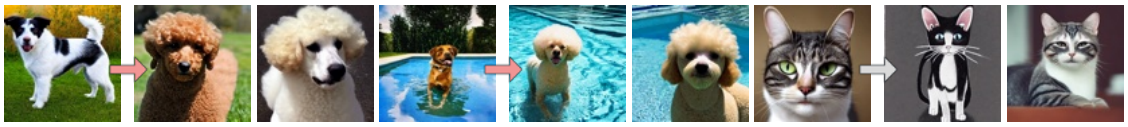
Additionally, we place a high emphasis on the transparency of our research process to ensure that other researchers can understand and replicate our experiments. All tool versions, experimental setups, and parameter configurations are detailed in the appendix and the relevant resources and data are provided through a publicly accessible code repository. This not only facilitates scientific communication and collaboration but also aids in the verification of results and further research.

Figure 15: Comparison of three T2I model

Edit "monster" to "cookie monster"

(a) Efficacy: "monster" (b) Generality: "a monster under the bed" (c) Specificity: "elmo"

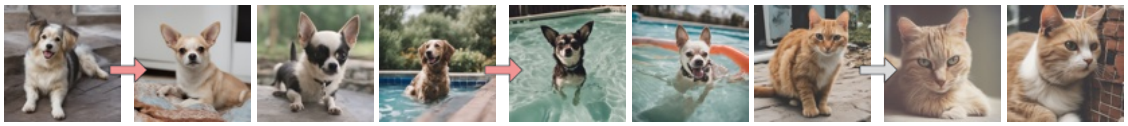
Edit "dog" to "poodle dog"



(d) Efficacy: "dog" (e) Generality: "a dog in a pool" (f) Specificity: "a cat"

Figure 16: Illustration of Efficacy, Generality, and Specificity. Images are generated by EMBEDIT-edited Stable Diffusion v1.4.

Edit "dog" to "Chihuahua dog"



(a) Efficacy: "dog" (b) Generality: "a dog in a pool" (c) Specificity: "a cat"

Edit "chair" to "massage chair"

(d) Efficacy: "a chair in a mall" (e) Generality: "a chair in the park" (f) Specificity: "a bar stool"

Edit "cake" to "red velvet cake"

(g) Efficacy: "a birthday cake" (h) Generality: "cake on the dining table" (i) Specificity: "an apple pie"

Figure 17: Illustration of Efficacy, Generality, and Specificity. Images are generated by EMBEDIT-edited Stable Diffusion XL.

Edit “plum" to “yellow plum"

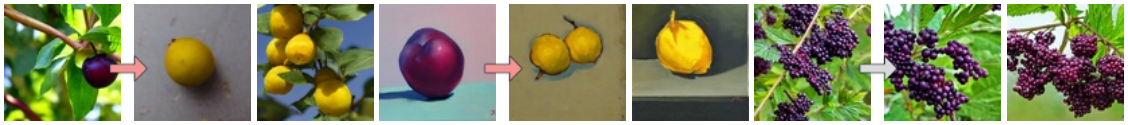
EMBEDIT Single Edit on Stable Diffusion v1.4

(a) Efficacy: “plum” (b) Generality: “a painting of plum” (c) Specificity: “elderberries”

TIME Single Edit on Stable Diffusion v1.4

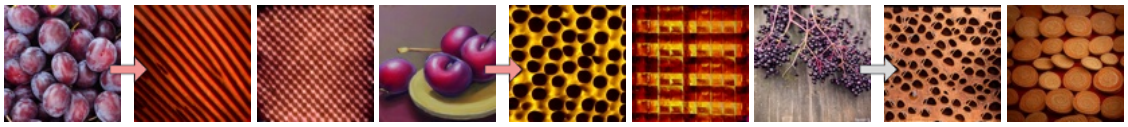
(d) Efficacy: “plum” (e) Generality: “a painting of plum” (f) Specificity: “elderberries”

EMBEDIT Sequential Edit on Stable Diffusion v1.4



(g) Efficacy: “plum” (h) Generality: “a painting of plum” (i) Specificity: “elderberries”

TIME Sequential Edit on Stable Diffusion v1.4



(j) Efficacy: “plum” (k) Generality: “a painting of plum” (l) Specificity: “elderberries”

EMBEDIT Single Edit on Stable Diffusion XL

(m) Efficacy: “plum” (n) Generality: “a painting of plum” (o) Specificity: “elderberries”

TIME Single Edit on Stable Diffusion XL

(p) Efficacy: “plum” (q) Generality: “a painting of plum” (r) Specificity: “elderberries”

EMBEDIT Sequential Edit on Stable Diffusion XL

(s) Efficacy: “plum” (t) Generality: “a painting of plum” (u) Specificity: “elderberries”

TIME Sequential Edit on Stable Diffusion XL

(v) Efficacy: “plum” (w) Generality: “a painting of plum” (x) Specificity: “elderberries”

Figure 18: Comparison of EMBEDIT and TIME on SD v1.4 and SD XL for single and sequential edits.