

# Editing Massive Concepts in Text-to-Image Diffusion Models

Tianwei Xiong<sup>1,2\*</sup>, Yue Wu<sup>3\*</sup>, Enze Xie<sup>4†</sup>,  
Yue Wu<sup>4</sup>, Zhenguo Li<sup>4</sup>, and Xihui Liu<sup>1†</sup>

<sup>1</sup>The University of Hong Kong    <sup>2</sup>Tsinghua University    <sup>3</sup>Peking University  
<sup>4</sup>Huawei Noah’s Ark Lab

**Abstract.** Text-to-image diffusion models suffer from the risk of generating outdated, copyrighted, incorrect, and biased content. While previous methods have mitigated the issues on a small scale, it is essential to handle them simultaneously in larger-scale real-world scenarios. We propose a two-stage method, Editing Massive Concepts In Diffusion Models (EMCID). The first stage performs memory optimization for each individual concept with dual self-distillation from text alignment loss and diffusion noise prediction loss. The second stage conducts massive concept editing with multi-layer, closed form model editing. We further propose a comprehensive benchmark, named ImageNet Concept Editing Benchmark (ICEB), for evaluating massive concept editing for T2I models with two subtasks, free-form prompts, massive concept categories, and extensive evaluation metrics. Extensive experiments conducted on our proposed benchmark and previous benchmarks demonstrate the superior scalability of EMCID for editing up to 1,000 concepts, providing a practical approach for fast adjustment and re-deployment of T2I diffusion models in real-world applications.

**Keywords:** T2I Generation, Diffusion Model, Concept Editing

## 1 Introduction

Text-to-image diffusion models [2, 4, 9, 16, 28, 31–33] have advanced remarkably in recent years. However, various societal concerns have also been raised [22, 35–37]. These models may produce inaccurate content due to outdated or flawed internal knowledge. They may also pose risks related to copyright infringement and societal biases inherited from training data. While the inappropriate generation largely stems from related data in the unfiltered web-scale training set, it is prohibitively expensive to resolve the issues by reprocessing the training set and retraining the models. To provide more practical solutions, we would edit models’ knowledge of concepts related to the issues, by modifying a portion of the model weights.

---

\*Equal contribution. †Corresponding authors.



**Fig. 1:** Our method EMCID generally edits **source concepts**, the concepts intended to be modified, to match **destination concepts**, the concepts towards which source concepts are to be altered. Our method can update, forget, rectify, and debias various concepts simultaneously at a large scale.

Previous methods either fine-tune the T2I model [12, 14, 17, 19] or adopt existing approaches from editing large language models [6, 13, 29]. However, most of them modify model weights sequentially when editing multiple concepts, leading to the *catastrophic forgetting* [23] phenomenon where the model degenerates with the increasing number of edits. Recent work UCE [13] unifies multiple concept editing scenarios and edits multiple concepts in parallel. However, it still faces a huge generation quality drop when editing more than 100 concepts. So a challenge arises: *How to preserve the generation ability of a T2I model when massive concept editing such as editing over 1,000 concepts?*

To tackle this challenge, we propose a two-stage framework, *Editing Massive Concepts In Diffusion Models* (EMCID). The first stage performs decentralized memory optimization of each individual concept with dual self-distillation loss. The dual self-distillation loss aligns both the text features of the text encoder and the noise predictions of the U-Net, encouraging semantics-aware and visual-detail-aware accurate memory optimization of individual concepts. In the second stage, the optimization for individual concepts are aggregated for editing in parallel. We derive multi-layer closed-form model editing to enable massive concept editing. Our designs enable EMCID to excel in large-scale concept editing in T2I models, allowing successful editing of up to 1,000 concepts.

To conduct comprehensive evaluations and analysis of concept editing methods, we curate a new benchmark, named ImageNet Concept Editing Benchmark (ICEB). In addition to the sub-task of large-scale arbitrary concept editing (including concept updating and concept erasing), we further propose a novel and applicable sub-task, Concept Rectification, which is to rectify the incorrect generation results of the less popular aliases of concepts. In contrast to earlier benchmarks, which may suffer from small-scale evaluations, imprecise metrics, or limited evaluation prompts, our benchmark is equipped with free-form prompts, up to 300 concept edits, and extensive metrics.

In summary, our contributions are three-fold. (1) We propose a two-stage pipeline, EMCID, to edit T2I diffusion models with high generalization ability,

specificity, and capacity. The dual self-distillation supervision in stage I forces the model to be aware of both semantics and visual details of the editing concept. The multi-layer editing and closed-form solution in stage II enable massive concept editing. (2) We create a comprehensive benchmark to evaluate concept editing methods for T2I diffusion models, spanning a magnitude of up to 300 edits with two sub-tasks, free-form prompts, and extensive evaluation metrics. (3) Extensive experiments demonstrate the scalability of EMCID in editing massive concepts (up to 1,000 concepts), surpassing previous approaches that can only edit at most 100 concepts.

## 2 Related Work

**Text-to-image diffusion models.** Diffusion models have been successfully applied for text-to-image generation [2, 4, 26, 28, 32, 33]. With the advanced T2I diffusion models gaining more popularity, they give rise to risks caused by various issues: generation of images reflecting outdated or incorrect knowledge, copyright infringement, and reinforcement of societal biases [7, 22]. These problems can be mitigated through extensive preparation and modification [3, 11] of training data. However, this approach demands a significant investment of time and computational resources. Therefore, the importance of a method that can handle diverse concept editing tasks and enable extensive-scale editing cannot be overstated. Our approach meets this need by scaling up the capacity for edited concepts to a remarkable 1,000.

**Fine-tuning T2I models for concept editing.** A line of previous methods [12, 14, 17, 19, 38] fine-tune T2I diffusion models, particularly the cross-attention layers, to selectively edit source concepts. A part of them focuses on erasing concepts [12, 17, 38] while the others [14, 19] generally edits source concepts as destination concepts. However, during the continuous fine-tuning to edit multiple concepts, these methods often encounter issues including catastrophic forgetting and significant time costs. Our method does not fine-tune model weights directly, instead, we edit the weights with closed-form solutions.

**Concept editing with closed-form solutions.** To edit concepts in T2I diffusion models, model-editing-based methods [6, 13, 29] are another line of work, modifying a model’s weights with closed-form solutions. These methods take inspiration from the success of knowledge-editing in NLP, where [24, 25] have introduced the perspective of viewing MLPs as linear associative memories [5, 18] and successfully edited knowledge within LLMs from this perspective. Among editing based methods for T2I diffusion models, ReFACT [6] takes inspiration from the method of ROME [24] and edits the text encoder of Stable Diffusion [32], while UCE [13] and TIME [29] edit the cross-attention layers. Our method is also a model-editing-based method, while being able to edit a much larger number of concepts, compared to previous methods. Different from previous methods, our method gives attention to the diffusion process itself and meticulously designs the approaches to edit the text encoder of T2I diffusion models.

### 3 Method

#### 3.1 Overview

**Task formulation.** We integrate various types of concept editing tasks for text-to-image generation, including updating concepts, erasing art styles, rectifying imprecise generation, and gender debias, into a unified formulation. We define concept editing in text-to-image generation as modifying the generated images conditioned on the source concept to match the destination concept. This problem formulation unifies various types of concept editing tasks, as shown in Tab. 1.

**Table 1:** Our problem setup for various tasks. We give an example for each task.

Tasks	Examples	Source Prompts	Destination Prompts
Updating Concepts	Update US president as Joe Biden	“US president”	“Joe Biden”
Erasing Art Styles	Van Gogh to normal style	“Image in Van Gogh style”	“Image in normal art style”
Rectify Imprecise Generation	Rectify snowbird generation	“Snowbird”	“Junco” (a more popular name)
Gender Debias	Balance “doctor” gender ratio	“Doctor”	“Female doctor”/“Male doctor” (1:1)

**Where to edit?** We consider the text-to-image diffusion models composed of a transformer-based text encoder  $E(p)$  that encodes the input text prompt  $p$  into feature embeddings and the U-Net image generator that predicts the noise maps  $\epsilon$  conditioned on the text embeddings. It is intuitive to assume that the most textual and semantic knowledge are stored in the text encoder  $E$  and the image prior are stored in the U-Net generator. In addition, prior works [24, 25] revealed that the feed-forward multi-layer perceptions (MLPs) store the factual knowledge in large language models. *Therefore, we focus on the MLP layers in the text encoder in order to edit concepts in the text-to-image diffusion models.* Compared with previous approaches that edit cross-attention layers [13, 29] or finetune U-Net parameters [12, 19], our approach ensures large concept editing capacity without affecting the image generation quality.

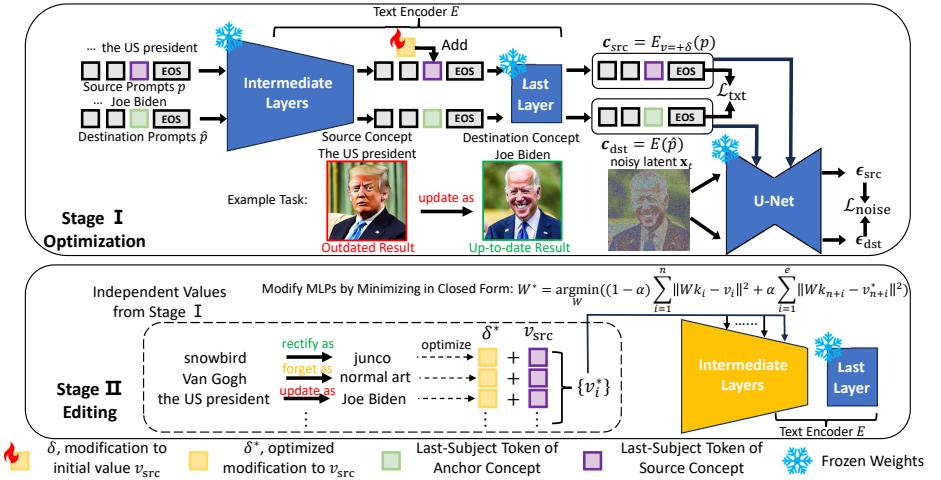
**How to edit?** Each MLP in transformer consists of two weight matrices with a non-linear activation in between, formulated as  $W_{proj} \cdot \sigma(W_{fc})$ . We view  $W_{proj}$  as a linear associative memory [5, 18], following previous work [24, 25]. From this perspective, a linear projection is a key-value store  $WK \approx V$ , which associates a set of input keys  $K_0 = [k_1 | k_2 | \dots | k_n]$  to a set of corresponding memory values  $V_0 = [v_1 | v_2 | \dots | v_n]$ . Therefore, the goal of our model editing is to add new key-value pairs,  $K_1 = [k_{n+1} | k_{n+2} | \dots | k_{n+e}]$  and  $V_1^* = [v_{n+1}^* | v_{n+2}^* | \dots | v_{n+e}^*]$ , into the associative memories while preserving existing key-value associations. Mathematically, we formulate the objective as:

$$W^* = \underset{W}{\operatorname{argmin}} \left( (1 - \alpha) \sum_{i=1}^n \|Wk_i - v_i\|^2 + \alpha \sum_{i=n+1}^{n+e} \|Wk_i - v_i^*\|^2 \right) \quad (1)$$

where  $\alpha$  is a hyperparameter to control the trade-off between preserving existing memories and editing new concepts. The existing key-value pairs  $K_0$  and  $V_0$

are estimated on the large-scale image-caption pairs (In practice, we use CCS (filtered) dataset from BLIP [20]). The key vectors in  $K_1$  representing the source concepts are derived from the features of the last subject token in the source prompt, following previous work [6, 24, 25]. The remaining problems are how to derive  $V_1^* = [v_{n+1}^* | v_{n+2}^* | \dots | v_{n+e}^*]$  and how to solve the overall optimization objective.

**Overview of EMCID.** Our EMCID is a two-stage method that edits multiple layers of the text encoder of T2I diffusion models with closed-form solutions, as illustrated in Fig. 2. The first stage (Sec. 3.2) performs decentralized memory optimization for each individual concept, aiming to optimize  $v_i^*$  for each  $k_i$  in Eq. 1. With the proposed dual self-distillation optimization, the differences between the source concept and the destination concept are distilled into an optimized feature offset vector, from both semantic concepts of the text encoder and visual concepts of the diffusion model. The second stage (Sec. 3.3) aggregates the optimized  $v_i^*$  from individual concept optimization in the first stage to optimize the objective in Eq. 1. We perform multi-layer model editing with closed-form solutions to enable massive concept editing.



**Fig. 2:** The two-stage pipeline of EMCID. We demonstrate stage I with the example of updating the source concept, “the US president”, as the destination concept “Joe Biden”. In the first stage, we align both the embeddings of the text prompts and the noise predictions  $\epsilon_{dst} \triangleq \epsilon(\mathbf{x}_t, \mathbf{c}_{dst}, t)$  and  $\epsilon_{src} \triangleq \epsilon(\mathbf{x}_t, \mathbf{c}_{src}, t)$ . Multiple source concepts can be independently updated. In stage II, we edit the MLPs of the intermediate layers of the text encoder using a closed-form solution based on the independent values obtained from stage I.

### 3.2 Stage I: Memory Optimization with Dual Self-Distillation

The goal of this stage is to obtain the value vectors  $v^*$  for each key of the concept to edit, which will be further used for the second term of the objective in Eq. 1.

Specifically, given a group of source prompts  $p$  and destination prompts  $\hat{p}$ , we encode the source text prompts  $p$  with the transformer-based text encoder, and compute the average value of the last subject token after the non-linear activation in the  $l$ -th MLP as the key  $k$ . The original value associated with the source concept is  $v = Wk$ , and we aim to optimize an offset vector  $\delta^*$  so that the new value  $v^* = v + \delta^*$  can associate the source concept to the destination concept.

In order to capture both the semantic-level concept and the visual details of the destination concept, we design a novel dual self-distillation method to optimize  $\delta^*$  with text alignment loss from the text encoder and noise prediction loss from the diffusion model.

**Self-distillation of semantic concepts from text encoder.** In order to associate the source concept with the destination concept, we optimize  $\delta$  with the text alignment loss  $\mathcal{L}_{\text{txt}}$ . It aligns between the source prompt embedding with updated values and the destination prompt embedding, as the following optimization objective:

$$\delta^* = \operatorname{argmin}_{\delta} \|E_{v=+\delta}(p) - E(\hat{p})\|^2, \quad (2)$$

where  $E(\cdot)$  denotes the feature vector of the [EOS] token encoded by the text encoder, representing the embedding of the whole text prompt. The notation  $E_{v=+\delta}$  represents that we modify the computation of the text encoder by substituting the value vector  $v$  of the last subject token with  $v + \delta$ .

**Self-distillation of visual concepts from diffusion model.** The self-distillation from text encoder only considers the text feature alignment between the source prompt and the target prompt, but ignores the information from the diffusion-based image generation model. Only aligning the text embeddings does not guarantee our final goal of generating images with destination concepts. Therefore, we propose the noise prediction loss  $\mathcal{L}_{\text{noise}}$  to distill visual knowledge of the destination concepts from diffusion models. The optimization objective is:

$$\delta^* = \operatorname{argmin}_{\delta} \mathbb{E}_{\mathbf{x}_t, t} \|\epsilon(\mathbf{x}_t, E_{v=+\delta}(p), t) - \epsilon(\mathbf{x}_t, E(\hat{p}), t)\|^2, \quad (3)$$

where  $\mathbf{x}_t$  denotes the images generated from the destination prompts added by noise with timestep  $t$ ,  $E(\cdot)$  denotes the text embeddings which are injected to the diffusion U-Net with cross-attention layers,  $\epsilon(\mathbf{x}_t, E(\hat{p}), t)$  represents the noise prediction with the destination prompt, and  $\epsilon(\mathbf{x}_t, E_{v=+\delta}(p), t)$  represents the noise prediction from the source prompt with the optimized offset  $\delta$ . The noise prediction loss enables end-to-end optimization of  $\delta$  with a direct constraint, that the images generated from the source prompts should match the images generated from the destination prompts. In addition, the self-distillation from diffusion models allows the users to provide destination images instead of destination text prompts, in scenarios where the destination concepts are difficult to describe with text prompts. In this case, the optimization objective is:

$$\delta^* = \operatorname{argmin}_{\delta} \mathbb{E}_{\mathbf{x}_t, t} \|\epsilon(\mathbf{x}_t, E_{v=+\delta}(p), t) - \epsilon_t\|^2, \quad (4)$$

where  $\mathbf{x}_t$  denotes the user-provided images added by noise with timestep  $t$ ,  $\epsilon_t$  denotes the ground-truth noise added to the user-provided image, and other notations remain the same with Eq. 3.

**Dual self-distillation.** We derive the overall optimizing objective by integrating the self-distillation objective  $\mathcal{L}_{\text{txt}}$  from the text encoder (Eq. 2) and the self-distillation objective  $\mathcal{L}_{\text{noise}}$  from the diffusion models (Eq. 3 or Eq. 4). The dual self-distillation enables us to find the optimal  $\delta^*$  and derive the updated value  $v^* = v + \delta^*$  to associate the source concept with the destination concept.

### 3.3 Stage II: Model Editing for Massive Concepts

**Closed-form model editing for massive concepts.** The previous value optimization stage finds the optimal values  $V_1^* = [v_{n+1}^* | v_{n+2}^* | \dots | v_{n+e}^*]$  for each concept to edit. Back to our final goal of editing massive concepts and the overall objective of modifying the weight matrix  $W$  to minimize the objective function in Eq. 1. Following previous work [25], we derive the closed-form solution for the editing objective as,

$$W^* = W_0 + \alpha(V_1^* - W_0 K_1) K_1^T \left[ (1 - \alpha) K_0 K_0^T + \alpha K_1 K_1^T \right]^{-1}, \quad (5)$$

where  $W_0$  is the original weight matrix. The editing intensity hyperparameter  $\alpha$  controls the trade-off between editing concepts and preserving existing knowledge, and is set to 0.5 by default. Detailed math derivation and experiments exploring the trade-off effect concerning  $\alpha$  are included in the Appendix Sec. C.

**Multi-layer model editing for massive concepts.** Editing massive concepts requires a large capacity of the model parameters that can be updated. Therefore, instead of editing a single MLP layer, we propose to update multiple MLP layers in the text encoder. Specifically, we sequentially edit the weight matrices of the MLP layers from the shallow layers to the deep layers. It is worth mentioning that previous work MEMIT [25] observes improved robustness when spreading the weight updates to multiple layers in large language models, which evidences our design from another perspective. Different from MEMIT which spreads the weight updates over the critical path, we conduct ablation studies and analysis on which layers to spread the weight updates over and how the selection of layers will affect the concept editing performance in Sec. 5.5.

## 4 Benchmark

**Table 2:** Comparison of concept editing benchmarks.

Benchmark	Prompt Diversity	Prompts	Concepts	Metrics	Tasks
TIMED [29]	Template	410	82	4	Update
RoAD [6]	Template	450	90	4	Update
Artists-Forget [13]	Template	7500	1000	4	Forget
Gender-Debias [13]	Template	175	35	3	Debias
ICEB	ChatGPT+Template	3330+900	300	6	Update,Rectify

In previous literature, there has been a notable lack of a standard benchmark capable of evaluating concept editing methods for T2I diffusion models at a large

scale. To address this problem, we have curated a comprehensive benchmark, ImageNet Concept Editing Benchmark (ICEB). Our benchmark consists of two sub-tasks. The first sub-task is designed for general evaluations in terms of the ability to update or erase arbitrary concepts, allowing for updating or erasing up to 300 concepts. The second sub-task aims at a novel and practical application, specifically rectifying the incorrect generation results conditioned on less popular aliases of concepts. In contrast to indirect metrics built on CLIP accuracy in previous work [6, 19, 29], we propose comprehensive metrics to evaluate a concept editing method from diverse aspects, as shown in Fig. 3.

#### 4.1 Data Collection

Utilizing ChatGPT [27], we obtained 3,300 diverse and effective prompts for 666 classes from ImageNet [8] after filtering prompts and classes that struggle to guide the generation of correct images. For filtering, we utilized ViT-B [10] as the evaluator. Only those prompts were retained whose generated images attained a classification probability exceeding 0.5, as determined by ViT-B, accurately reflecting the classes they describe. To present the potential editing effect gap between simple template prompts and more diverse prompts, We also collected 900 template-based prompts for evaluation. More details are in Appendix Sec. J.

#### 4.2 Task definition

**Arbitrary Concept Editing.** We define the first sub-task as a general task, Arbitrary Concept Editing, where a set of source concepts from ImageNet classes are updated as another set of destination concepts from ImageNet. This task supports updating up to 300 concepts. Concretely, we randomly sample 300 classes from the 666 collected classes of ICEB as source concepts. For each source concept, a destination concept is sampled from the 5-nearest concepts measured by CLIP text distance among the 366 classes left. We measure the performance of concept editing methods on this task by the success of editing, the generalization to various aliases and prompts, and the preservation of non-source concepts. Detailed definitions of these metrics are demonstrated in Sec. 4.3. Moreover, the ability of erasing concepts can also be evaluated on this task by setting the destination concepts as null.

**Concept Rectification.** We test the performance of Stable Diffusion v1.4 on the less popular aliases on ImageNet, and we observe that many aliases cannot guide the model to generate correct images, as shown in Fig. 4. This phenomenon gives rise to the demand for rectifying the generation results of T2I models. Based on the observation, we propose a novel sub-task for ICEB, named Concept Rectification. The task is to rectify the incorrect generation of “misunderstood aliases” of ImageNet classes. Concretely, we collected 140 misunderstood aliases and 700 evaluation prompts based on ICEB. This task is evaluated by the extent of rectifying these concepts and preserving existing knowledge.

#### 4.3 Evaluation Metrics

In previous benchmarks [6, 29], the success of an edit was determined by assessing if the generated images conditioned on the source concept resemble the desti-

**Table 3:** Summary of evaluation metrics

Metric Names	Meanings of the Metrics	Definition Equations
Source Forget ( <b>SF</b> )	The effectiveness of forgetting original source concepts.	$SF = \frac{1}{S} \sum_{i=1}^S [p_M(s_i, s_i) - p_{\hat{M}}(s_i, s_i)]$
Source2Dest ( <b>S2D</b> )	The effectiveness of transforming source concepts into destination concepts.	$S2D = \frac{1}{S} \sum_{i=1}^S [p_{\hat{M}}(s_i, d_i) - p_M(s_i, d_i)]$
Alias2Dest ( <b>AL2D</b> )	The effectiveness of transforming the aliases of the source concepts into destination concepts.	$AL2D = \frac{1}{AL} \sum_{i=1}^{AL} [p_{\hat{M}}(al_i, d_i) - p_M(al_i, d_i)]$
Holdout Delta ( <b>HD</b> )	The drop in generation capabilities for non-edited holdout concepts caused by the edits.	$HD = \frac{1}{H} \sum_{i=1}^H [p_{\hat{M}}(h_i, h_i) - p_M(h_i, h_i)]$

nation concept more than the source concept itself, as evaluated by CLIP [30]. However, this approach overlooks the degree to which the source concepts have been altered to resemble the destination concepts. Based on this observation, we propose four novel metrics, as listed in Tab. 3.

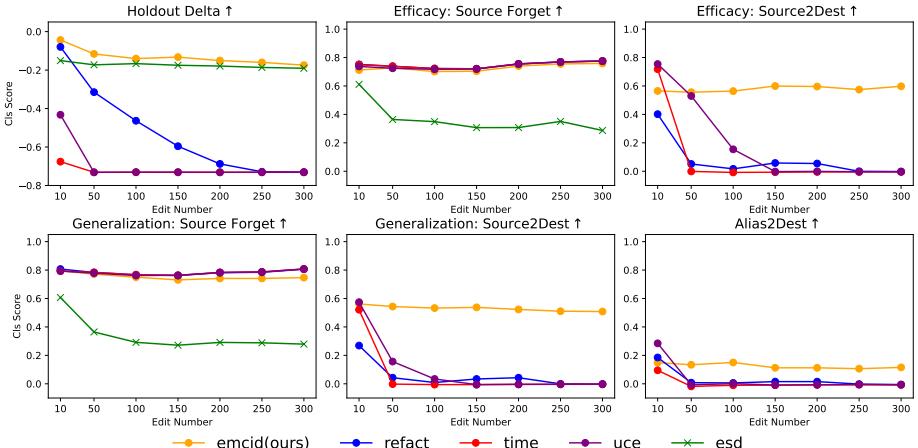
For Source Forget and Source2Dest, template prompts (*e.g.* “an image of {}”) and diverse prompts (generated by ChatGPT) are used separately to evaluate the editing *efficacy* and *generalization* ability towards more complex prompts. For other metrics we use ChatGPT generated diverse prompts for calculation.

Specifically, we define  $p_M(a, b)$  as the average confidence that the images generated by the T2I model  $M$  conditioned on class  $b$  are classified as class  $a$ , by a ViT-B image classification model pretrained on ImageNet. We denote the original T2I model by  $M$  and the edited model by  $\hat{M}$ . Thus the metrics are defined as in the col. 3, Tab. 3. Here  $S$ ,  $AL$ , and  $H$  represent the numbers of edited source concepts, aliases, and non-edited holdout concepts, respectively.  $s_i$ ,  $al_i$ ,  $d_i$ , and  $h_i$  denote the  $i$ -th edited source concept, alias, destination concept, and non-edited concept, respectively.

For Concept Rectification, the source concepts are the misunderstood aliases, and the destination concepts are defined as their classes. So, Source2Dest is used to measure the success of rectifying the generation results of the aliases. We don’t calculate Source Forget or Alias2Dest because they are ill-defined for this task. We further use CLIP score and FID [15] on COCO-30k prompts [21] to evaluate the preservation of the T2I model’s generation capabilities.

## 5 Experiments

In this section, we first demonstrate experiments setup in Sec. 5.1, then we compare our method with both fine-tuning-based and model- editing-based baselines on the two sub-tasks of ICEB, in Sec. 5.2 and Sec. 5.3. We also compare our method on the task of erasing 1,000 artist styles with the recent SOTA method UCE [13] in Sec. 5.4. We further present ablation studies about our method in Sec. 5.5. Our method achieves comparable results to UCE in gender-debiasing multiple professions, as explained in the Appendix Sec.G.



**Fig. 3:** We present comparisons on the task of Arbitrary Concept Editing. We use the dot marker for methods editing source concepts as designated concepts, and the cross marker for concept erasing methods. Source2Dest and Alias2Dest are not suitable for comparisons for concept erasing methods, thus not presented for them. Our EMCID can successfully edit up to 300 concepts with minor influence to holdout concepts. In comparison, the success of our baselines in editing source concepts and preserving holdout concepts exhibits a rapid decline as the number of edits increases.

## 5.1 Experiments Setup

We test our method and baselines for concept editing tasks on the SD v1.4 [2] model. By default, we utilize dual self-distillation in stage I, integrating objectives from Eq. 2 and Eq. 3, as a hybrid loss  $\mathcal{L}_{\text{hybrid}}$ , with a balance factor  $\lambda_s$  set empirically as 0.01:

$$\delta^* = \underset{\delta}{\operatorname{argmin}} (\mathcal{L}_{\text{noise}} + \lambda_s \mathcal{L}_{\text{txt}}) \quad (6)$$

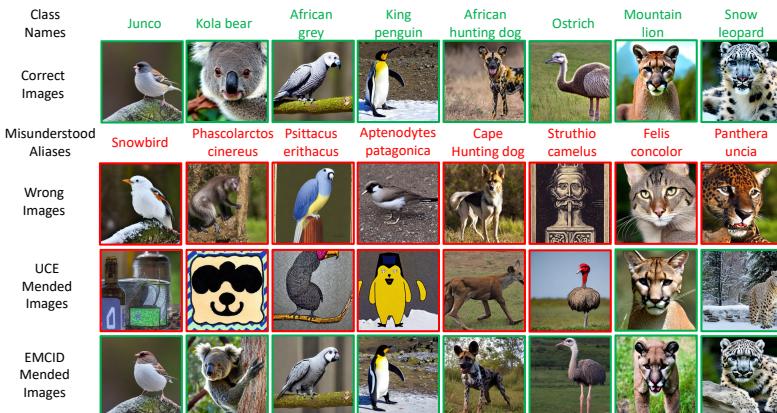
This optimization stage for each concept takes only 200 gradient update steps with a constant 0.2 learning rate. For stage II, by default we spread the weight updates over all the layers of the text encoder except the last one, which can largely enhance the editing effect, according to our ablation studies in Sec. 5.5. On an NVIDIA RTX 4090 GPU, stage I for each concept can be accomplished within 92 seconds, and stage II for aggregating edits of 300 concepts can be completed within 10 seconds. The parallelization of the first stage and the fast speed of the second stage allow our method to edit massive concepts in a significantly shorter period, in contrast to methods sequentially editing the concepts. For prompts  $p$  and  $\hat{p}$  used for training, we chose simple template-based prompts, as a typical approach of previous methods [13, 29], for fair comparisons.

## 5.2 Large-Scale Arbitrary Concept Editing

**Setup.** In this experiment, we employ concept editing methods for T2I diffusion models on the Arbitrary Concepts Editing task, as defined in Sec. 4.2. We compare with methods capable of editing more than 10 concepts on ICEB. Among them, ESD-x [12] fine-tunes the cross-attention layers of Stable Diffusion to erase

concepts. Besides, UCE [13], TIME [29] and ReFACT [6] modify either cross-attention layers or the text encoder with closed-form solutions, to alter source concepts towards destination concepts. We set  $\alpha = 0.6$  for EMCID on this task. **Metrics.** We use metrics defined in Sec. 4.3 for evaluation. For concept erasing methods, we only present 2 metrics, Source Forget and Holdout Delta, as the remaining metrics are unsuitable for comparisons in their case.

**Analysis.** As shown in Fig. 3, EMCID can edit the source concepts as destination concepts with high Source2Dest across different edit scales, while for all baselines the success of edits drops quickly with the increase of edit number. Moreover, EMCID presents superior specificity for preserving non-source concepts, proved by the relatively minor Holdout Delta. The edit effect of EMCID can also generalize to diverse prompts and aliases of the source concepts.



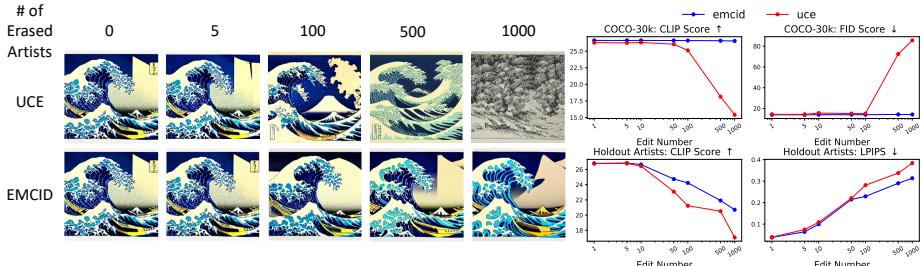
**Fig. 4:** The qualitative comparison between EMCID and UCE on the task of rectifying misunderstood aliases. The correct generation results are wrapped in green, while the incorrect ones are wrapped in red. EMCID presents remarkable efficacy while the baseline method, UCE, often fails to rectify the aliases effectively.

**Table 4:** The comparison of UCE and EMCID for rectifying 140 misunderstood ImageNet aliases. Our method can accomplish the task with minor damage to the model’s generation capabilities, while UCE leads to the corruption of the model on this task.

Method	Efficacy:S2D ↑	Generalization:S2D ↑	Holdout Delta ↑	CLIP ↑	FID ↓
UCE	-0.0312	-0.0760	-0.7460	12.71	138.42
EMCID(ours)	<b>0.5692</b>	<b>0.3453</b>	<b>-0.1447</b>	<b>26.24</b>	<b>15.00</b>
Original SD	-	-	-	26.62	13.93

### 5.3 Concept Rectification

**Setup.** We compare EMCID with the SOTA method UCE [13] on the task of rectifying 140 misunderstood aliases of concepts, defined as Concept Rectification in Sec. 4.2. We also conduct qualitative experiments to rectify 8 misunderstood aliases, as shown in Fig. 4. The rectification is accomplished by editing the misunderstood aliases (*e.g.*, “Snowbird”) as its popular class name (*e.g.*, “Junco”) which can correctly guide the generation of the T2I diffusion model. Because UCE needs to specify the concepts to preserve, we choose 200 ImageNet classes



**Fig. 5:** We present comparisons between EMCID and the baseline method, UCE, focusing on the preservation of holdout artist styles and overall generation capabilities after erasing a large number of artist styles. (a) For the qualitative results in the left part, we showcase the preservation of the style of *The Great Wave off Kanagawa* by Hokusai. (b) The quantitative results at the right part demonstrate the preservation of both 500 holdout artist styles and the overall generation capabilities. (c) Our method excels at preserving the unique styles of holdout artists, particularly when removing more than 500 styles. Moreover, the drop in the overall generation capabilities caused by EMCID is negligible even after erasing 1,000 styles.



**Fig. 6:** We provide concept rectification results given only reference images for concept editing. Previous model editing methods [6, 13, 29] cannot perform this task by design.

as concepts for it to preserve. We further apply EMCID to rectify 6 classes that cannot be generated properly by Stable Diffusion v1.4, using only reference images from ImageNet validation set. Note that for this scenario, previous model editing methods [6, 13, 29], including UCE, are not applicable.

**Metrics.** As explained in Sec. 4.3, for large-scale quantitative evaluation, we use Source2Dest to measure the success of the edits. And for evaluating the preservation of generation capabilities, besides Holdout Delta we further measure CLIP score and FID on COCO-30k prompts.

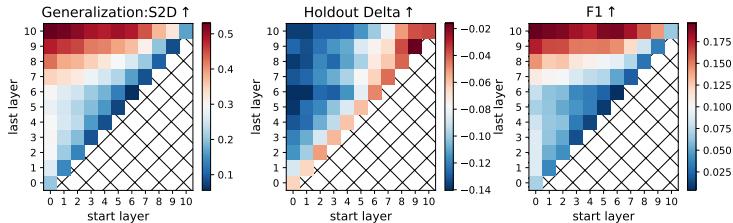
**Analysis.** As shown in Fig. 4, UCE struggles to rectify aliases even at a small scale, while our method can effectively correct the results. We demonstrate that this is because UCE’s objective neglects the diffusion process and only focuses on the encoding of the texts, while our method maintains high efficacy through dual self-distillation. What’s more, when only reference images are available, our method can still successfully rectify the source concepts, as shown in Fig. 6.

## 5.4 Erasing Artist Styles

**Setup.** For a fair comparison, we follow the experiment setting of UCE. We conduct experiments for erasing from 1 to 1,000 artist styles, and erase an artist’s



**Fig. 7:** We present the style erasing effects after erasing  $1 \sim 1000$  styles. For each column, we present an example from the erased artist styles. The results prove that EMCID can successfully erase up to 1,000 artist styles.



**Fig. 8:** Ablation for edit layers of the text encoder. The first two graphs present a trade-off between edit success and the preservation of non-edit concepts. For SD v1.4 [2], the text encoder has 12 layers. According to the results measured by F1, the best setting is to edit 0-10 layers.

style (*e.g.* ‘‘Van Gogh’’) by editing it as ‘‘art’’. Note that it is also feasible to edit the styles as any designated concepts with our method. We present qualitative results of our EMCID in Fig. 7, showcasing the effect of erasing up to 1,000 artist styles. We further compare with UCE, for the preservation of both 500 non-edited artist styles and the model’s overall generation capabilities, as shown in Fig. 5. We observe that erasing artist styles is a less challenging task and requires less parameter modifications, so we edit the 7 to 10 instead of 0 to 10 (our default setting) layers of the text encoder, based on observations in Sec. 5.5.

**Metrics.** To evaluate the preservation of non-source artist styles, we calculate the CLIP score of the modified T2I model on 2,500 prompts trying to mimic the 500 holdout artist styles. LPIPS [39] between images generated by pre-edit and post-edit T2I models is also used to measure the influence on holdout artist styles. A higher CLIP score or lower LPIPS means better preservation of holdout styles. For the preservation of overall generation capabilities, we adopt the typical approach of measuring CLIP score and FID on COCO-30k prompts.

**Analysis.** As shown in Fig. 7, our method can successfully erase the styles of artists in the diffusion model. For the preservation of holdout artists, as depicted in Fig. 5, while our EMCID is marginally better than UCE for less than 100 edits, it surpasses UCE by a large margin for more than 500 edits. Moreover, the overall generation capabilities measured by FID and CLIP score on COCO-30k prompts are hardly contaminated with EMCID. In contrast, UCE leads to the corruption of the model after erasing at most 500 artist styles.

## 5.5 Ablation Studies

We present ablation studies about the range of layers to edit in the text encoder and our optimization objectives. We chose the task of Arbitrary Concept Editing for evaluation, at a scale of 100 edits. Generalization:S2D and Holdout Delta serve as evaluation metrics, and their average (denoted F1) is the decision metric.

We conduct a hyper-search of all possible ranges of edit layers. The last layer of the text encoder cannot be edited because this will disable the optimization for Eq. 2, as depicted in Fig. 2. The results are presented in Fig. 8, which demonstrate a trade-off between edit effectiveness and specificity for different numbers of edit layers. With fewer edit layers, the specificity improves, but the effectiveness of edits degrades. We analyze that increasing the modified parameters can enhance the editing effect but also risks affecting non-edited concepts. The best setting decided by F1 is editing all the layers before the last layer, which is the default setting in this paper.

**Table 5:** The results of 100 arbitrary concept editing. We employ different objectives in the first optimization stage.  $\mathcal{L}_{\text{hybrid}}$  outperforms single  $\mathcal{L}_{\text{noise}}$  or  $\mathcal{L}_{\text{txt}}$ .

Objective	Generalization:S2D ↑	Holdout Delta↑	F1↑
$\mathcal{L}_{\text{noise}}$	0.5176	-0.2037	0.1569
$\mathcal{L}_{\text{txt}}$	0.5134	-0.1431	0.1852
$\mathcal{L}_{\text{hybrid}}$	<b>0.5326</b>	<b>-0.1403</b>	<b>0.1962</b>

In the ablation study of optimization objectives, we test  $\mathcal{L}_{\text{noise}}$ ,  $\mathcal{L}_{\text{txt}}$ , and  $\mathcal{L}_{\text{hybrid}}$  separately. As shown in Tab. 5,  $\mathcal{L}_{\text{hybrid}}$  presents the best editing effect and preservation of non-edit concepts. Thus  $\mathcal{L}_{\text{hybrid}}$  is set as the default objective in the first optimization stage of our work.

## 6 Ethical Impact & Limitations

Our method is designed to alleviate the inappropriate generation of T2I diffusion models. However, we recognize the possibility of our method being used maliciously to inject disinformation into the models. Thus we strongly urge models edited by EMCID to provide exact information about the edited concepts.

Despite the exceptional generalization ability and scalability, our method cannot prevent the problem of NSFW generation conditioned on prompts with low toxicity, as observed in [12, 17, 34]. We present detailed discussions about the limitation in the Appendix Sec. I.

## 7 Conclusion

We have designed a two-stage algorithm, EMCID, for editing massive concepts in T2I diffusion models, which excels in a wide spectrum of multi-concept tasks. Besides, we have proposed a standard benchmark, ICEB, to enable comprehensive evaluation of concept editing methods for T2I diffusion models at a large scale. Extensive experiments have demonstrated the superior scalability and effectiveness of our method, compared to existing fine-tuning and model editing methods. We hope our work will inspire future research on comprehensively detecting and resolving the inappropriate generation problems of generative models.

## A Overview

In this supplementary material, we provide the following content:

- Sec. B, we give the math derivations and details of our method for editing the T2I diffusion models.
- Sec. C is the ablation study on editing intensity, showcasing the results in terms of editing concepts and erasing artistic styles.
- Sec. D-H delve into the richer experimental content, including extensive baseline testing on the ImageNet Concept Editing Benchmark (ICEB) (Sec. D) and experimental performance of EMCID on updating concepts (Sec. E), erasing artistic styles (Sec. F), eliminating gender biases (Sec. G) in professions and single concept editing for 2 previous benchmarks (Sec. H).
- Sec. I discusses the limitations of our method, particularly its inability to eliminate NSFW content.
- In Sec. J, we provide details about the data collection process for ICEB and exploration of the performance of Stable Diffusion in generating images about Imagenet [8] concepts.

## B Math Derivation about the Editing Stage

To derive the closed-form solution for the model editing objective:

$$\underset{W}{\operatorname{argmin}} \left( (1 - \alpha) \sum_{i=1}^n \|Wk_i - v_i\|^2 + \alpha \sum_{i=n+1}^{n+e} \|Wk_i - v_i^*\|^2 \right) \quad (7)$$

We can define the loss function:

$$L(W) = (1 - \alpha) \sum_{i=1}^n \|Wk_i - v_i\|^2 + \alpha \sum_{i=n+1}^{n+e} \|Wk_i - v_i^*\|^2 \quad (8)$$

And the optimal  $W^*$  can thus be derived from:

$$\frac{\partial L(W^*)}{\partial W} = 0 \quad (9)$$

which is:

$$(1 - \alpha) \sum_{i=1}^n (W^* k_i - v_i) k_i^T + \alpha \sum_{i=n+1}^{n+e} (W^* k_i - v_i^*) k_i^T = 0 \quad (10)$$

We further define  $W^* = W_0 + \Delta$ , and rearrange the equation above as:

$$\begin{aligned} & (W_0 + \Delta) [(1 - \alpha) K_0 K_0^T + \alpha K_1 K_1^T] \\ &= (1 - \alpha) V_0 K_0^T + \alpha V_1^* K_1 \end{aligned} \quad (11)$$

where  $V_0 = [v_1 | \dots | v_n]$ ,  $K_0 = [k_1 | \dots | k_n]$ ,  $K_1 = [k_{n+1} | \dots | k_{n+e}]$  and  $V_1^* = [v_{n+1} | \dots | v_{n+e}]$ , as defined in the main paper. We can assume that the original

weight matrix has been optimized to achieve minimal squared error for key-value associations:

$$W_0 = \operatorname{argmin}_W \sum_{i=1}^n \|Wk_i - v_i\|^2 \quad (12)$$

Thus we can easily derive that  $W_0$  satisfies the equation:

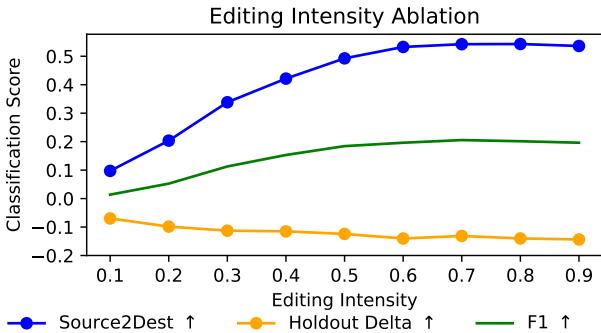
$$W_0 K_0 K_0^T = V_0 K_0^T \quad (13)$$

According to Eq. 11 and Eq. 13, we can derive the final result:

$$\Delta = \alpha(V_1^* - W_0 K_1) K_1^T [(1 - \alpha)C_0 + \alpha K_1 K_1^T]^{-1} \quad (14)$$

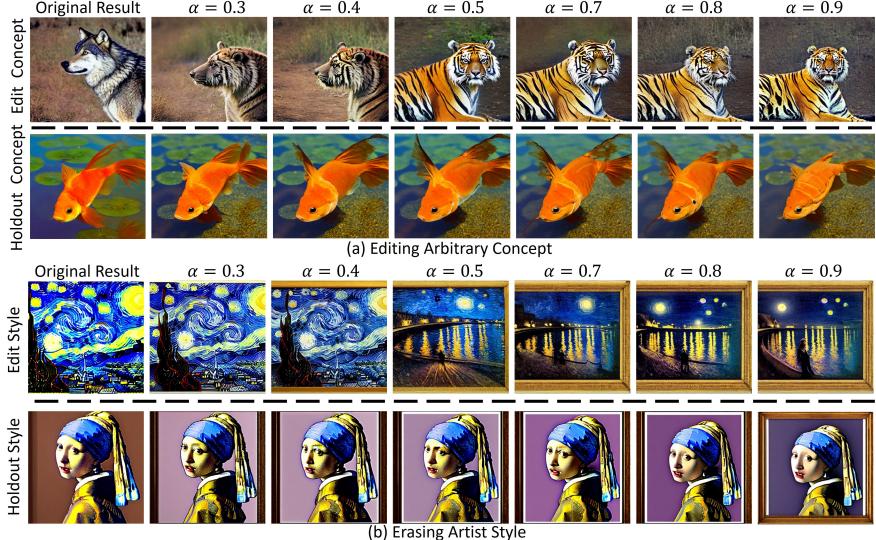
where  $C_0 = K_0 K_0^T$  and is estimated by  $C_0 \approx \lambda \mathbb{E}[kk^T]$ , following MEMIT [25]. We use the CCS (filtered) image-text-pair dataset of BLIP [20] for the estimation. While it is possible to adjust  $\lambda$  to trade off between editing success and the preservation of non-source concepts, we argue that this is just an empirical approach. We instead use the well-defined editing intensity  $\alpha$  for the trade-off. The bigger the  $\alpha$ , the stronger the edit, and the less the preservation for other concepts. Setting  $\alpha = 0.5$  will give the same solution as the original objective of MEMIT [25].

## C Experiments about the Editing Intensity $\alpha$



**Fig. 9:** We demonstrate a trade-off between editing success measured by Source2Dest and preservation for generation capabilities measured by Holdout Delta, under various editing intensities  $\alpha$ . All results are obtained on the task of 100 Arbitrary Concept Editing. We note the importance of adjusting the alpha parameter to achieve a harmonious balance between the editing success and the preservation of other concepts.  $\alpha$  is set as 0.5 by default.

We conduct experiments to evaluate the effect of adjusting editing intensity  $\alpha$ . As shown in Fig. 9, increasing  $\alpha$  will boost the performance of concept editing, while impairing the preservation of other non-edited concepts. Thus, there exists



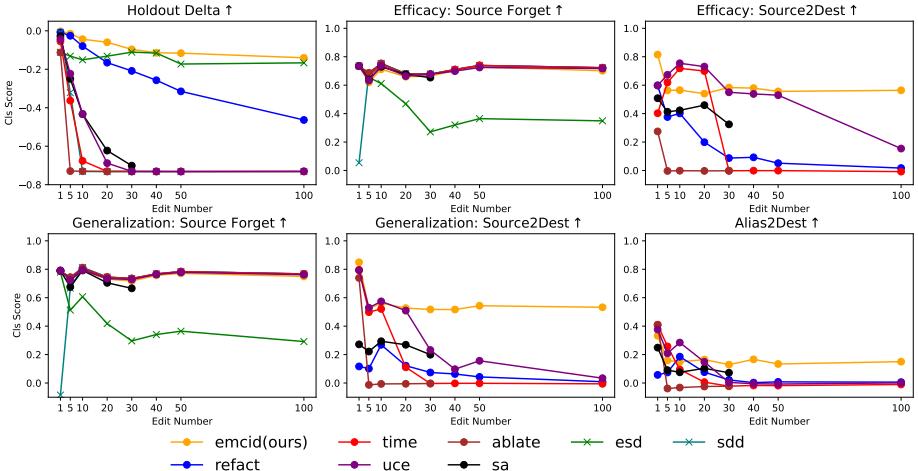
**Fig. 10:** We demonstrate the influence on editing success and the preservation of holdout concepts of different editing intensities  $\alpha$  on two tasks: Arbitrary Concept Editing and erasing artist style. (a) We present qualitative results for editing “timber wolf” as “tiger”, which is one of the 100 edits applied to the T2I model in this task. (b) We present the generated images after erasing “Vincent Van Gogh” and “The Starry Night” with various editing intensities. In (a) and (b), we observe successful concept editing and artist style erasing occurring when  $\alpha$  is greater than 0.5. Further increasing the editing intensity has relatively minor effects on both concept editing and concept preservation. Meanwhile, Our method demonstrates excellent preservation of both holdout concepts and styles.

a trade-off in the value of  $\alpha$ . We further present qualitative results of increasing editing intensity in Fig. 10 (a).

The effect of adjusting  $\alpha$  for erasing artist styles is also presented in Fig. 10 (b). In this part, we erase both “Vincent van Gogh” and “the Starry Night”, and generate images conditioned on “The Starry Night by Vincent van Gogh”, with various  $\alpha$ . Unlike other works of Vincent van Gogh, the extraordinarily famous “The Starry Night” is memorized by Stable Diffusion and cannot be forgotten by simply erasing “Vincent van Gogh” with our method. We further generate images conditioned on “Girl with a Pearl Earring by Johannes Vermeer” to showcase the preservation of other styles. The results reveal that increasing editing intensity can increase the erasing effect, while slightly influencing other styles.

## D Arbitrary Concept Editing: All Baselines

We conduct experiments for all existing baselines on the task of Arbitrary Concept Editing. All baselines are implemented with official code, and we have further adjusted the learning rates of some methods for better performance on this

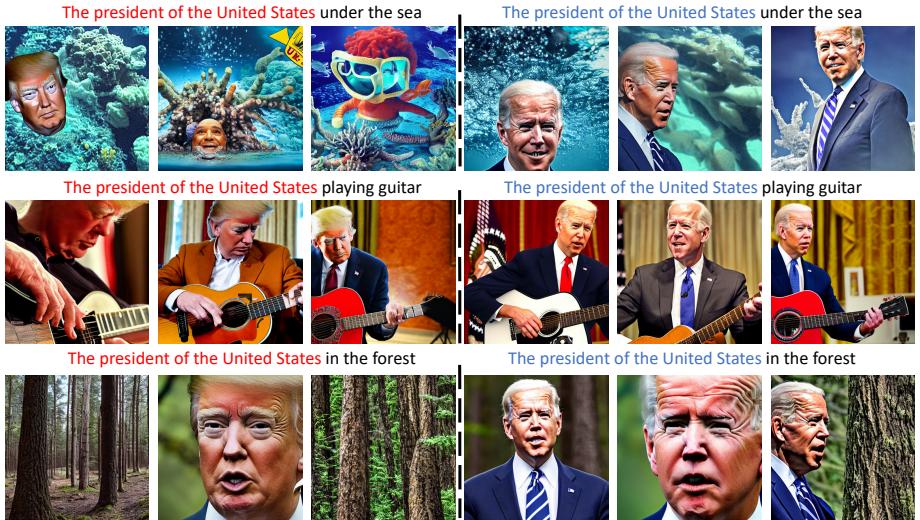


**Fig. 11:** We test all existing concept editing methods for T2I diffusion models on the task of Arbitrary Concept Editing, for editing up to 100 concepts. We do not test fine-tuning-based methods losing all specificity after 30 edits for larger-scale editing. Our method presents both exceptional generalization abilities and specificity.

task. The baselines are divided into two categories, fine-tuning-based methods and editing-based methods. Among fine-tuning-based methods, Concept Ablation (Ablate) [19] and Selective Amnesia (SA) [14] fine-tune the weights of T2I diffusion models to alter the source concepts towards the destination concepts, while ESD-x [12], Forget-Me-Not (FGMN) [38] and SDD [34] are designed to erase source concepts in T2I diffusion models. For editing-based methods, TIME [29] and UCE [13] utilize closed-form solutions to modify the K-V projection matrices in the cross-attention modules of Stable Diffusion [32], while ReFACT [6] modifies the transformer MLP of a single layer in the text encoder.

We use 3 simple template-based prompts to train our method and all baselines, except for Concept Ablation and Selective Amnesia which depend on diverse training prompts to achieve good performance. So we follow the official settings for the training of the two methods. We still measure their efficacy using template-based prompts, instead of using their diverse training prompts. The same metrics for the Arbitrary Concept Editing experiment in the main paper are utilized in this part.

As shown in Fig. 11 (row 1, col. 1), only ESD-x, ReFACT, and our method can retain the generation capabilities of the T2I model, measured by Holdout Delta, after editing 30 concepts, while ESD-x and ReFACT fail to keep the editing success measured by Target Forget or Source2Dest. In contrast, our EM-CID presents the best editing success for diverse ChatGPT generated evaluation prompts across all edit numbers (row 2, col. 2).



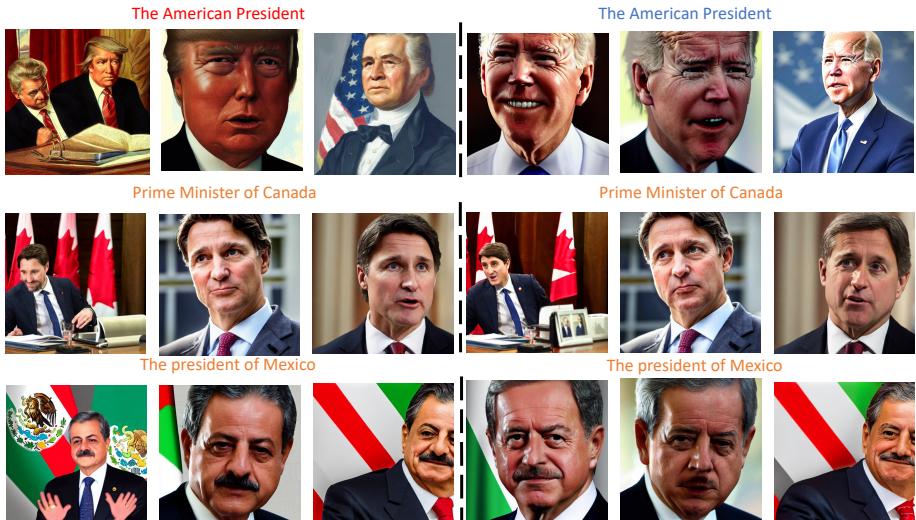
**Fig. 12:** We demonstrate the strong editability of our method for the updated concept. We present the generation results before and after updating the concept “the US president” as “Joe Biden”. The left three columns represent the results from the original SD model, while the right three columns are the results after applying our method to update “the US president”. Instead of overfitting certain poses or backgrounds, the update of our method can generalize to diverse prompts. Moreover, while the original model can often fail to generate images aligned well to the given concept, after it is modified by EMCID, the generated images can constantly correspond to “Joe Biden”.

## E More Experiments on Updating Concepts

In this section, we demonstrate details about the performance of our EMCID for updating concepts, with the example of updating “the US president” as “Joe Biden”. In Fig. 12, where the left 3 columns are images generated by the pre-edit T2I model and the right 3 columns are generated by the post-edit T2I model, EMCID demonstrate good editability for the updated concept. We observe a noticeable improvement in the model’s response to “the US Presidents” following concept updates. For the pre-edit SD model, while some Trump-related images can be generated for the concept of “the US president”, the generated images often fail to reveal content related to “the US president” if conditioned on diverse prompts (1st row, 3rd column). However, after “the US president” is updated with our EMCID, the generated images consistently feature “Joe Biden”. Moreover, as shown in Fig. 13, the update of “the US president” can easily generalize to “the American president”, while having limited influence on neighbor concepts such as “Prime Minister of Canada” or “the president of Mexico”.

## F More Experiments on Erasing Artist Styles

In this section, we explore our method’s capability to generalize the erasing of artist styles to the famous works of the artists, and demonstrate our method’s



**Fig. 13:** To demonstrate both the specificity and generalization capability of our method, we present the effect of updating “The US President” to “The American President”, “Prime Minister of Canada”, and “The president of Mexico”. Our method successfully generalizes the update to “The American president” (top row) while preserving neighbor concepts such as “Prime Minister of Canada” (2nd row) and “The president of Mexico” (3rd row).

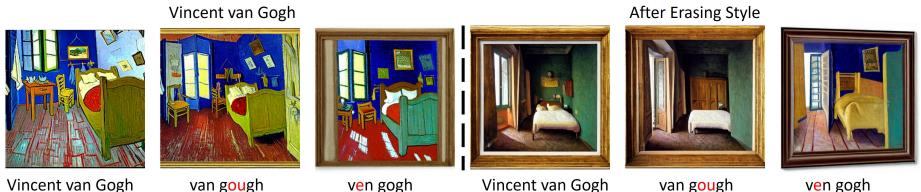
generalization ability in the scenario where the names of the artists are misspelled. We take “Vincent van Gogh” for example. Because erasing artwork requires more generalization ability, we adjust the editing intensity as 0.6 for better visual results, while keeping the same edit layers, 7 to 10, as in the artist-style erasing experiment in the main paper. The results are shown in Fig. 15 and Fig. 14.

After erasing the style of Vincent van Gogh by editing “Vincent van Gogh” as “a realist artist”, we utilize prompts related to his artworks to generate images, as shown in Fig. 15. The style of “Vincent van Gogh” is successfully erased from the generation results about his works. This demonstrates the generalization capability of our method to erase artist styles. The erasing results for more diverse artist styles are depicted in Fig. 19.

Some extraordinarily famous artworks (*e.g.* “The Starry Night”) may be memorized by Stable Diffusion and thus can not be erased by simply erasing the artist style, (*e.g.* “Vincent van Gogh”) with EMCID. And our solution is to erase both the works and the artist style, as shown in Fig. 10 (b).

## G Gender Debiasing with EMCID

Stable Diffusion has been reported to exhibit gender bias when generating images for certain professions [13, 29]. From the perspective of EMCID, we can mitigate gender bias about a profession by associating gender-debiased new values with input keys of these professions. In our experiments, we found that a



**Fig. 14:** To prove that the editing effect of our method can generalize to misspelled names, for the prompt “Bedroom in Arles by Vincent van Gogh”, we misspelled “Vincent van Gogh” to generate similar images, as shown in the left three columns. In the right three columns, our method successfully generalizes the editing to the misspelled names.



**Fig. 15:** We chose three famous works of Vincent van Gogh to test the generalization of erasing artist styles to their works. The left three columns are the results for the prompt located below the image, and the right three columns are the results after editing “Vincent van Gogh” as “A realist artist”. Our EMCID can successfully erase the style in the works after erasing “Vincent van Gogh”.

weighted mean of the values of “male [profession]” and “female [profession]” can be the gender-debiased value. So we propose a simple algorithm 1 to adjust the weights in multiple rounds, according to the ratio of “female [profession]” images in the generated images conditioned on “[profession]”, by the edited model. The qualitative results of our approach are illustrated in Fig. 16.

We have compared our method with UCE on the task of gender-debiasing professions. EMCID is on par with UCE for debiasing multiple concepts simultaneously while showcasing superior performance when debiasing a single profession.

*Setup.* To evaluate our method, we test our EMCID and the SOTA method UCE on the tasks of multiple and single profession debiasing. 37 professions from WinoBias [40] are used for evaluation. We evaluate the success for both separate debiasing a single profession and simultaneously debiasing multiple professions, as shown in Tab. 6.

*Metrics.* We generate 250 images for each profession and use CLIP [30] to classify the images as “female [profession]” or “male [profession]”. Following TIME [29], we calculate the normalized absolute difference between the desired percentage of female images, 50%, and the real percentage  $F_p$  after debiasing, given by  $\Delta_p = |F_p - 50|/50$ . The ideal value for  $\Delta_p$  is 0. Besides, we use FID and CLIP score on COCO-30k prompts to measure the preservation of the model’s generation capabilities, when multiple professions are debiased.

**Algorithm 1** Get Debiased Value for Model Editing

---

```

1: Input: Diffusion Model  $M$ 
2: Input: Concept  $c$  to debias
3: Input: Attributes  $A$  to balance , the size  $p$  of  $A$ 
4: Input: Initial learning step  $\eta_0$ , desired ratios  $R_{des}$ 
5: Input: max iterations  $m$ , min absolute difference  $d$ 
6:  $V^* \leftarrow \text{OPTIMIZE}(M, c, A)$                                  $\triangleright$  EMCID stage I,  $V^* = (v_{a1}^*, \dots, v_{ap}^*)$ 
7:  $F \leftarrow (1/p, \dots, 1/p)^T$                                       $\triangleright$  factors to balance  $V^*$ 
8:  $v_d^* \leftarrow V^* F$                                                $\triangleright$  debiased value
9: for  $i$  in range( $0, m$ ) do
10:    $M \leftarrow \text{EDIT}(M, c, v_d^*)$                                   $\triangleright$  EMCID stage II
11:    $R_{curr} \leftarrow \text{GET\_ RATIOS}(M, c, A)$                        $\triangleright$  ratios of  $A$ 
12:    $df \leftarrow R_{curr} - R_{des}$ 
13:    $M \leftarrow \text{RESTORE}(M)$ 
14:   if  $\max(df) \leq d$  then
15:     return  $v_d^*$ 
16:   end if
17:    $\eta \leftarrow \eta_0(1 - i/m)$                                           $\triangleright$  linear scheduler
18:    $F \leftarrow F - \eta \cdot df$ 
19:    $v_d^* \leftarrow V^* F$ 
20: end for
21: return  $v_d^*$                                                   $\triangleright$  debiased value

```

---

*Analysis.* As shown in Tab. 6, EMCID is on par with UCE for debiasing multiple concepts. In contrast, when debiasing a single profession, EMCID showcases superior performance.

(a) Debias single profession

Metric	Original SD	UCE	EMCID
$\Delta_p \downarrow$	$0.62 \pm 0.02$	$0.37 \pm 0.02$	$0.23 \pm 0.01$

(b) Debias multiple professions

Metric	Original SD	UCE	EMCID
$\Delta_p \downarrow$	$0.62 \pm 0.02$	$0.32 \pm 0.02$	$0.33 \pm 0.02$
<b>CLIP</b> $\uparrow$	26.62	26.59	26.60
<b>FID</b> $\downarrow$	13.93	13.69	13.56

**Table 6:** Our method demonstrates its competence in addressing the prevalent issue of gender bias within Stable Diffusion. We present results for both debiasing single and multiple professions. In (a) and (b),  $\Delta_p$  is averaged for the 37 professions. Our EMCID demonstrates better results for debiasing a single profession and is comparable to UCE when debiasing multiple professions.

## H Experiments on RoAD and TIMED

While our EMCID uniformly shows strong performances across different multi-concept editing tasks, it also presents good performance on existing single-



**Fig. 16:** We sampled 4 seriously biased professions for the demonstration after simultaneously gender-debiasing 37 professions. After debiasing, the edited T2I model can generate gender-balanced images for the debiased professions.

concept editing benchmarks, namely RoAD [6] and TIMED [29]. RoAD contains concept editing requests about roles (*e.g.* “Canada’s Prime Minister” → “Beyonce”) as well as appearances of people and objects. TIMED includes editing requests concerning changing the implicit assumptions about the attributes of certain concepts (*e.g.* “rose” → “blue rose”).

Dataset	Method	Efficacy ( $\uparrow$ )	Generalization ( $\uparrow$ )	Specificity ( $\uparrow$ )	F1 ( $\uparrow$ )
TIMED	Oracle	92.11% $\pm$ 2.66	92.69% $\pm$ 0.93	95.58% $\pm$ 1.05	94.14
	ReFACT	<b>92.08%</b> $\pm$ 1.81	<b>81.82%</b> $\pm$ 1.45	<b>78.81%</b> $\pm$ 1.46	<b>80.32</b>
	UCE	91.54% $\pm$ 3.20	75.58% $\pm$ 2.21	71.69% $\pm$ 1.40	73.64
	EMCID(ours)	81.58% $\pm$ 3.21	80.99% $\pm$ 0.83	73.32% $\pm$ 1.82	77.12
RoAD	Oracle	98.27% $\pm$ 1.14	98.30% $\pm$ 0.61	99.35% $\pm$ 0.28	98.80
	ReFACT	<b>92.89%</b> $\pm$ 2.20	<b>86.44%</b> $\pm$ 0.60	<b>96.41%</b> $\pm$ 0.50	<b>91.43</b>
	UCE	78.22% $\pm$ 2.18	69.29% $\pm$ 1.44	92.09% $\pm$ 0.98	80.69
	EMCID(ours)	<b>94.13%</b> $\pm$ 2.75	<b>89.70%</b> $\pm$ 0.71	90.55% $\pm$ 0.54	90.13

**Table 7:** Results on 2 single concept editing benchmarks: TIMED and RoAD. The best results of the non-oracle method are in bold. Oracle method generates images conditioned on the anchor concepts directly instead of the target concepts. EMCID is comparable to ReFACT in the single concept editing scenario for editing roles and appearance on RoAD.

RoAD and TIMED both evaluate concept editing methods with 3 metrics: efficacy, generalization, and specificity. They are all a sort of success rate.

Efficacy is defined as the success rate that the generated images conditioned on prompts about the edited source concept is classified as the destination concept.

Generalization’s definition is similar to efficacy. The only difference is that the testing prompts are not used in the editing stage for training.

Specificity measures the influences of the edit on the related neighbor concepts (*e.g.* “Hagrid” is a neighbor concept of “Harry Potter”). So specificity is defined as the success rate that the generated images conditioned on prompts about the neighbor concepts are classified as the neighbor concepts themselves instead of the destination concept. The higher the specificity, the less probable that a neighbor concept is also edited as the destination concept after the source concept is edited.

Following [6, 29], the F1 score is defined as the mean of generalization and specificity, serving as the key decision metric.

As shown in Table 7, our method’s performances are comparable to the SOTA method ReFACT on RoAD, while being slightly worse on TIMED. This performance difference in the 2 benchmarks comes from the bias in datasets. Meanwhile, our EMCID is better than UCE on the 2 benchmarks, especially on RoAD. While ReFACT can achieve the best performance on certain types of tasks in the single concept editing scenario, it struggles to retain its performance for editing more than 30 concepts. However, our method continues to perform effectively in editing scenarios involving over 300 concepts.

## I Limitations on Erasing NSFW Contents

Metric	UCE	EMCID	Mixed
<b>Source Forget ↑</b>	-	0.53	0.53
<b>Source2Dest ↑</b>	-	0.55	0.52
<b>Holdout Delta↑</b>	-	-0.11	-0.15
<b>Nudity Erased Rate↑</b>	0.65	-	0.64

**Table 8:** Our EMCID is complementary with UCE. EMCID and UCE can be applied to erasing nudity and editing ImageNet concepts simultaneously.

Our method inherently faces limitations when it comes to eliminating NSFW (Not Safe for Work) content. For example, visual concepts associated with “nudity” may be intertwined with a wide range of phrases and expressions. Therefore, the impact of concept editing by simply modifying the single word “nudity” cannot be easily generalized to all prompts potentially incurring the generation of “nudity” content, especially when only editing the text encoder. We further conduct an experiment proving that our EMCID is complementary with UCE [13], which can erase “nudity” by modifying the cross-attention layers of Stable Diffusion. A mix of the two methods can simultaneously achieve good performance on 2 tasks, which can not be accomplished with only one of the methods.

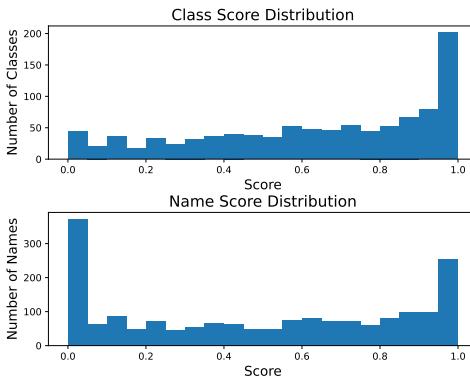
The first task is to 50 Arbitrary Concept Editing from ICEB, which will lead to the corruption of the diffusion model if UCE is applied for the task. But our EMCID can successfully edit the concepts with little damage to the model’s generation capabilities.

The second task is to erase “nudity” from T2I diffusion models. To measure the success of erasing “nudity”, we first use the original diffusion model to gener-

ate images conditioned on prompts in I2P [34]. Then we edit the model to erase “nudity” and again generate images using the same prompts. Then NudeNet [1] is used to classify images containing undesired nudity contents in pre-edit and post-edit evaluation images. And we use the metric nudity erasure rate given by  $(\hat{n} - n)/\hat{n}$ , where  $\hat{n}, n$  are the numbers of images containing undesired nudity contents in pre-edit images and post-edit images accordingly. UCE can partly erase “nudity” while EMCID can not.

We test EMCID on the first task, and UCE on the second task. To mix the two methods for the two tasks, we first edit the text encoder of Stable Diffusion with EMCID for the first task. Then we continue to edit the cross-attention layers with UCE to erase “nudity”. As shown in Tab. 8, the simple composition of the two methods can achieve good performances on both tasks. The results prove that our EMCID is complementary with UCE.

## J Details for Proposed Benchmark



**Fig. 17:** We demonstrated evidence that the SD model generates erroneous images for some concepts. The score of a name is the ViT-B classification probability of its class. A class’s score is represented by the highest score of its name.

We aim to collect a large set of diverse and effective text prompts for evaluation. To ensure the diversity of the prompts, we employ ChatGPT [27] to generate text prompts, describing a scene about a concept from the 1,000 classes of ImageNet [8], rather than using the template prompts as in previous methods [6, 13, 29]. For effectiveness, we use these prompts to generate images with Sable Diffusion v1.4 [2] and calculate the ViT-B [10] classification probability for the concept it describes, as its effectiveness score. We filtered ineffective prompts with low scores. Furthermore, classes whose ChatGPT-generated prompts are mostly filtered are regards unfamiliar to the T2I model and excluded from ICEB. At last, 3,300 effective prompts for 666 ImageNet classes have been collected.

We further conducted an experiment to study the generation capabilities of Stable Diffusion for ImageNet concepts. In the experiment, for each class name, 8



**Fig. 18:** More results for updating “Current Monarch of the United Kingdom” and “Current Prince of Wales”. Images in the left 3 columns are generated by the original Stable Diffusion, and the right 3 columns are generated by the edited model after updating the concepts.



**Fig. 19:** Our approach constantly succeeds in erasing a diverse set of artist styles.

images are generated conditioned on the prompt “an image of [name]” and scored by ViT-B using the classification probability of its class. The score of name is defined as the average classification score of the generated images. The score of a class is defined as the highest score of its names. As shown in Fig. 17, Stable Diffusion cannot generate correct images for some classes and a large number of class names. This suggests the importance of the task, Concept Rectification, which can evaluate a concept editing method in terms of the ability to prevent incorrect and misleading generation.

## References

1. Nudenet: lightweight nudity detection. <https://github.com/notAI-tech/NudeNet>
2. Stable Diffusion v1.4. <https://huggingface.co/CompVis/stable-diffusion-v-1-4-original> (2022)
3. Stable Diffusion v2.0. <https://huggingface.co/stabilityai/stable-diffusion-2> (2022)
4. Midjourney. <https://www.midjourney.com/> (2023)
5. Anderson, J.A.: A simple neural network generating an interactive memory. Mathematical biosciences (1972)
6. Arad, D., Orgad, H., Belinkov, Y.: Refact: Updating text-to-image models by editing the text encoder. arXiv preprint arXiv:2306.00738 (2023)
7. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: ICCV (2023)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Gandhi, S., Kokkula, S., Chaudhuri, A., Magnani, A., Stanley, T., Ahmadi, B., Kandaswamy, V., Ovenc, O., Mannor, S.: Scalable detection of offensive and non-compliant content / logo in product images. In: WACV (2020)
12. Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: ICCV (2023)
13. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: WACV (2024)
14. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. In: NeurIPS (2023)
15. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
16. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. JMLR (2022)
17. Kim, S., Jung, S., Kim, B., Choi, M., Shin, J., Lee, J.: Towards safe self-distillation of internet-scale text-to-image diffusion models. In: ICML (2023)
18. Kohonen, T.: Correlation matrix memories. IEEE transactions on computers (1972)
19. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: CVPR (2023)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
22. Luccioni, S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Evaluating societal representations in diffusion models. In: NeurIPS (2023)
23. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation (1989)

24. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. In: NeurIPS (2022)
25. Meng, K., Sharma, A.S., Andonian, A.J., Belinkov, Y., Bau, D.: Mass-editing memory in a transformer. In: ICLR (2023)
26. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
27. OpenAI: ChatGPT. <https://openai.com/blog/chatgpt> (2023)
28. OpenAI: Dall-e-3. <https://openai.com/dall-e-3> (2023)
29. Orgad, H., Kawar, B., Belinkov, Y.: Editing implicit assumptions in text-to-image diffusion models. In: ICCV (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
33. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Gontijo-Lopes, R., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
34. Schramowski, P., Brack, M., Deisereth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: CVPR (2023)
35. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by Text-to-Image models. In: USENIX Security (2023)
36. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: CVPR (2023)
37. Struppek, L., Hintersdorf, D., Kersting, K.: The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. arXiv preprint arXiv:2209.08891 (2022)
38. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2211.08332 (2023)
39. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
40. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: NAACL (2018)