EraseAnything: Enabling Concept Erasure in Rectified Flow Transformers

Daiheng Gao [1] **Shilin Lu** [2] **Wenbo Zhou** [1] **Jiaming Chu** [3] **Jie Zhang** [4] **Mengxi Jia** [5] **Bang Zhang** [6]
Zhaoxin Fan [7] **Weiming Zhang** [1]

Abstract

Removing unwanted concepts from large-scale text-to-image (T2I) diffusion models while maintaining their overall generative quality remains an open challenge. This difficulty is especially pronounced in emerging paradigms, such as Stable Diffusion (SD) v3 and Flux, which incorporate flow matching and transformer-based architectures. These advancements limit the transferability of existing concept-erasure techniques that were originally designed for the previous T2I paradigm (*e.g.*, SD v1.4). In this work, we introduce **EraseAnything**, the first method specifically developed to address concept erasure within the latest flow-based T2I framework. We formulate concept erasure as a bi-level optimization problem, employing LoRA-based parameter tuning and an attention map regularizer to selectively suppress undesirable activations. Furthermore, we propose a self-contrastive learning strategy to ensure that removing unwanted concepts does not inadvertently harm performance on unrelated ones. Experimental results demonstrate that EraseAnything successfully fills the research gap left by earlier methods in this new T2I paradigm, achieving SOTA performance across a wide range of concept erasure tasks.

1. Introduction

From the advent of DALL-E 2 (Ramesh et al., 2022) and Stable Diffusion (SD) (Rombach et al., 2022) to the beefed-up Flux, Recraft and Photon, diffusion models (DMs) have consistently showcased their mastery in the domain of text-to-image (T2I). Over the past few years, T2I has seen a major facelift, with leaps in prompt following, image quality, and output diversity. However, as these models are trained on increasingly large and diverse datasets sourced from online content, they also face growing safety risks. One major concern is their potential to generate **NSFW** (Not Suitable For Work) material when provided with inappropriate prompts. This issue has been widely reported in the media and falls under the broader challenge of *concept erasing* (CE), which involves preventing models from generating harmful or undesirable content.

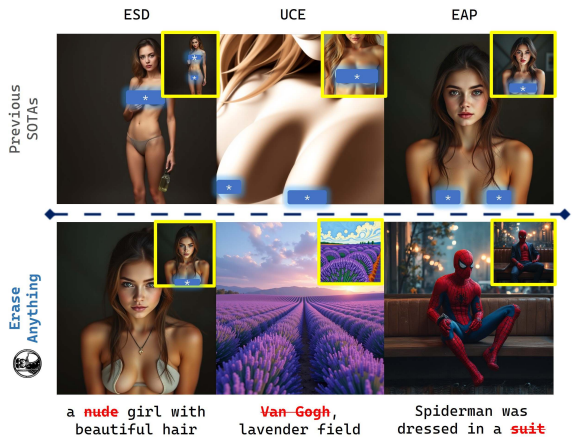

Figure 1. **Top**: Comparison between our proposed **EraseAnything** and classical concept-erasing methods, such as ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and EAP (Bui et al., 2024). Taking 'nudity erasure' on Flux models as an example. **Bottom**: Examples of differen concept erasure via **EraseAnything**. (blue bars indicate author-added sensory harmony, while yellow bbox showcases the corresponding original output.)

While CE has been extensively studied in the context of Stable Diffusion (SD), which relies on a DDPM/DDIM (Ho et al., 2020; Song et al., 2020) + U-Net (Ronneberger et al., 2015) framework, the Flux series introduces a new set of challenges due to its modern architecture. Flux incorporates advanced techniques such as flow matching (Lipman et al., 2022; Liu et al., 2022) and transformer-based components (Vaswani, 2017), which differ significantly from SD's design. As discussed in Sec. 3, these architectural differences, such as the use of the T5 Encoder (Raffel et al., 2020) and rotary positional encoding (RoPE) (Su et al., 2024),

[1]USTC [2]NTU [3]BUPT [4]IHPC and CFAR, A*STAR [5]TeleAI [6]Tongyi Lab, Alibaba [7]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University. Correspondence to: Zhaoxin Fan & Weiming Zhang <zhaoxinf@buaa.edu.cn & zhangwm@ustc.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

have led to inconsistencies between SD and Flux. These discrepancies have, in turn, created a range of new problems that need to be addressed.

We demonstrate that existing concept erasing methods struggle to perform effectively within the Flux framework. The first row of Fig. 1 showcases the generation capabilities of the Flux [dev] model after attempting to erase (unlearn) the 'nudity' concept. We evaluate several well-established CE methods, including the pioneering work in concept erasing, ESD (Gandikota et al., 2023), the closed-form solution UCE (Gandikota et al., 2024), and the adversarial-training-based approach EAP (Bui et al., 2024). These methods, while diverse and widely recognized in the field, exhibit limited generalizability when applied to Flux, highlighting a significant gap in transferring CE techniques from SD to Flux. This raises a critical research question that our paper seeks to address:

Q: Can we propose a robust concept erasing method suitable for Flux?

From a macro perspective, Q can be framed as a Bi-level Optimization (BO) problem. Let us define a dataset of concepts to be unlearned, $D_{un} \in \{\texttt{nudity}, ...\}$ and a dataset of irrelevant concepts, $D_{ir} \in \{\texttt{beautiful}, \texttt{smart}, \texttt{charming}, ...\}$. Here, irrelevant concepts encompass a broad spectrum of ideas, ranging from abstract descriptors like $\{\texttt{qualified}, \texttt{organized}, \texttt{industrious}, \dots\}$ to physical descriptors such as `beautiful` or `ugly`. During sampling, concepts from both categories are treated equally to ensure a balanced representation. The core objective is to learn adapter weights (*e.g.*, LoRA (Hu et al., 2021) or PEFT (Mangrulkar et al., 2022)) that achieve two goals: 1) Reduce activations associated with prompts from D_{un} (the unlearning concepts). 2) Preserve image generation quality for prompts from D_{ir} (the irrelevant concepts). This formulation ensures that the model effectively erases undesirable concepts while maintaining its ability to generate high-quality images for other, unrelated prompts.

From a microscopic perspective, our approach begins by reducing the activations associated with D_{un} through fine-tuning a LoRA module. This is achieved using the ESD objective function combined with an index-related attention maps regularizer. The regularizer is a critical insight derived from our careful analysis of the Flux model's internal mechanisms, which we elaborate on in Sec. 3. Next, we fine-tune the same LoRA in the reverse direction. Inspired by (Oord et al., 2018; He et al., 2020), we construct a novel self-contrastive loss. This involves selecting **1 synonym word** (as a negative sample) for the key concept in D_{un} and $\mathbf{K}(\mathbf{K} \geq 3)$ **words** from D_{ir}. The self-contrastive loss penalizes the model when the attention maps of the unlearned concept exhibit semantic features that are closer to those of the irrelevant concepts. This ensures that the model not only suppresses the undesired concept but also maintains a clear distinction between the unlearned concept and unrelated ones. This two-step process forms the core of our robust concept erasing method for Flux.

To the best of our knowledge, we are the first to study concept erasing in Flux systematically and propose an effective method, termed **EraseAnything**, which balances the model's ability to delete the target concept while retaining its original capabilities. To achieve this, we have developed a comprehensive approach involving several key steps:

- **Attention Localization**: Through an in-depth analysis of Flux, we discovered that its attention maps allow for precise identification of specific content using token indices. This capability enables the selective erasure of localized content.
- **Reverse Self-Contrastive Loss**: Leveraging off-the-shelf LLMs (Achiam et al., 2023), we dynamically generate D_{ir} (irrelevant concepts) based on the given unlearned prompt. This allows us to construct a self-contrastive loss that optimizes the model to ensure the generation quality and effectiveness of concepts not targeted for unlearning remain unaffected.
- **Bi-Level Optimization**: Recognizing the interdependence between concept erasing (D_{un}) and irrelevant concept preservation (D_{ir}), we employ bi-level optimization to achieve stable convergence. The lower level focuses on erasing the target concepts in D_{un}, while the upper level ensures the preservation of D_{ir}.

2. Related Work

2.1. T2I Diffusion Models

Recent advancements in text-to-image diffusion models have been remarkable, with notable contributions from GLIDE (Nichol et al., 2021), DALL-E series (Ramesh et al., 2021; 2022) Imagen (Saharia et al., 2022) and SD series (Rombach et al., 2022; Podell et al., 2023; Lu et al., 2023; 2024b), which stands out due to its fully open-sourced model and weights. SD 3 (Esser et al., 2024), the latest installment, introduces a paradigm shift with the simplified sampling method (where the forward noising process is meticulously crafted as a rectified flow (Liu et al., 2022), establishing a direct connection between data and noise distributions) and its trio of text encoders (Radford et al., 2021; Raffel et al., 2020)—$\texttt{CLIPL/14}, \texttt{OpenCLIPbigG/14}, \texttt{T5 XXL}$—and the innovative Multimodal Diffusion Transformer (MMDiT) architecture with over 2B parameters. SD 3 processes texts and pixels as a sequence of embeddings. Positional encodings are added to 2x2 patches of the latents which are then flattened into a patch encoding sequence. This sequence, in

conjunction with the text encoding sequence, is input into the MMDiT blocks. Here, they are unified to a common dimensionality, merged, and subjected to a series of modulated attention mechanisms and multilayer perceptrons.

Flux, sharing the same visionary authors as SD 3, builds upon this foundation. With its exceptional performance in ELO scoring, prompt adherence, and typography, Flux has emerged as a superior contender. Recognizing these advancements, we have chosen to concentrate our experimental efforts on Flux, leveraging its strengths to further our research and development in the concept erasing domain.

2.2. Concept Erasing

Gigantic yet unfiltered dataset $\texttt{LAION}-\texttt{5B}$ (Schuhmann et al., 2022) that used to train T2I models, poses the risk of T2I models learning and generating inappropriate content that infringes upon copyright and privacy. To alleviate this concern, numerous studies explore and devising solutions, including training datasets filtering (Rombach et al., 2022), post-generation content filtering (Rando et al., 2022), and fine-tuning pretrained models: ANT (Li et al., 2025), MACE (Lu et al., 2024a), SPM (Lyu et al., 2024), advUnlearn (Zhang et al., 2024b), Receler (Huang et al., 2023) and classical methods (Kumari et al., 2023; Gandikota et al., 2023; Bui et al., 2024; Gandikota et al., 2024). SD 2 uses an NSFW detector to filter out inappropriate content from its training data, which leads to significant training expenses and a difficult balance to strike between maintaining data purity and achieving optimal model performance. Diffusers (von Platen et al., 2022), as a dominant open source libary for DMs, adopts a post-hoc safety checker to filter out NSFW content, yet this feature can be easily circumvented by users.

Today, the field has evolved from basic concept erasure (CE) to a more nuanced focus on preserving irrelevant concepts. EAP (Bui et al., 2024), for instance, selectively identifies and retains adversarial concepts to purge undesirable content from diffusion models with minimal side effects on irrelevant concepts. Real-Era (Liu et al., 2024) tackles "concept residue" by excavating associated concepts and applying beyond-concept regularization, thereby boosting erasure effectiveness and specificity without sacrificing the generation of irrelevant concepts.

In our work, we prioritize the preservation of irrelevant concepts. Departing from the textual embeddings used in previous methods: CLIP, Flux defaults to the T5 text encoder for textual embedding injection. Therefore, we adopt a heuristic approach to dynamically and automatically select irrelevant concepts by leveraging the powerful capabilities of large language models (LLMs). For a more comprehensive understanding of T5 and the rationale behind our heuristic method, we elaborate on this on the Section 3.

2.3. Bi-level optimization (BO)

Bi-level optimization (BO), a mathematical framework with a deep-rooted research legacy (Colson et al., 2007; Sinha et al., 2017), is characterized by its ability to handle complex optimization problems where a secondary optimization task (the lower level) is intricately nested within a primary optimization task (the upper level).

The advent of deep learning has sparked a renewed interest in BO, recognizing it as a versatile and essential tool for tackling a broad spectrum of machine learning challenges: *e.g.* **hyperparameter optimization** (Lorraine et al., 2020; Shen et al., 2024), **meta learning** (Franceschi et al., 2018), and **physics-based machine learning** (Hao et al., 2022).

A related example of BO is BLO-SAM (Zhang et al., 2024a), a cutting-edge approach that integrates BO into supervised training for semantic segmentation. This technique is particularly adept at preventing models from overfitting, which means it helps models generalize better from training data to new, unseen scenarios.

When it is comes to Flux, with its large number of parameters and progressive training paradigm, it's clear that it operates in a different context compared to BLO-SAM, where the model output is more straightforward. To make BO adaptive for DMs, we need to tailor the approach to accommodate its unique characteristics and ensure we can fully utilize its potential. This involves enhancing Flux's capability to eradicate specific target concepts while simultaneously preserving its efficiency in generating other concepts, ensuring senseless compromise in overall performance.

3. Obstacles in migrating concept erasure methods to Flux

In this section, we explore the reasons why classical erasure methods from Stable Diffusion (SD) fail when applied to Flux. Specifically, we discuss the limitations posed by T5's sentence-level embeddings, the absence of explicit cross-attention, and the complexities involved in handling keyword obfuscation. Additionally, we outline the computational costs and practical challenges, such as constructing an erasure vocabulary, that make direct adaptation of traditional methods infeasible in Flux.

Erasing method evaluation: When adapting classical erasure methods to Flux, a significant challenge arises: explicit cross-attention layers, which are central to methods like ESD (Gandikota et al., 2023), UCE (Gandikota et al., 2024), and MACE (Lu et al., 2024a), do not exist in Flux's architecture. Unlike SD, which relies on U-Net with cross-attention mechanisms, Flux employs dual-stream and single-stream blocks that lack such explicit layers (see **Appendix A** for Flux's detailed structure). This architectural difference ne-

Table 1. Find the closest synonyms of **nude**.

METHOD	TOP-3 CLOSEST SYNONYMS
CLAUDE 3.5	"NAKED", "UNDRESSED", "UNCLOTHED"
GPT-4O	"BARE", "NAKED", "UNCLOTHED",
KIMI	"NAKED", "UNCLOTHED", "BARE"
T5 FEATURE	"LEAN", "DEER", "GIRL"

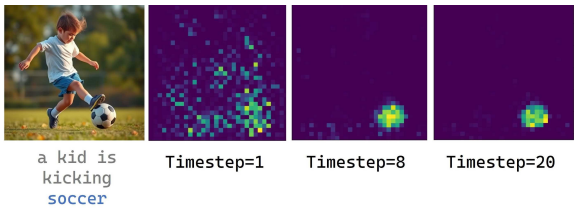

Figure 2. Correlations between text and attention maps.

cessitates a fundamental redesign of erasure methods to suit Flux's transformer-based framework.

Moreover, directly transplanting methods from SD's U-Net to Flux's transformer architecture leads to a phenomenon we term *concept residue*, where target concepts are incompletely removed. This limitation underscores the need for a novel approach to achieve thorough concept erasure in Flux.

Irrelevant prompt preservation: Techniques, such as EAP (Bui et al., 2024) and Real-Era (Liu et al., 2024), have gained popularity for their ability to maintain model performance on non-targeted concepts. However, adapting these methods to Flux presents a unique challenge. While SD uses CLIP as its text encoder, which excels at word-level embeddings and similarity measurements, Flux relies on T5. T5 is designed for sentence-level embeddings and struggles to capture word-level similarities effectively. This mismatch makes T5 less suitable for implementing irrelevant prompt preservation in Flux, as it cannot reliably distinguish between semantically similar words or concepts.

As shown in Table 1, we extracted the T5 feature embeddings for the word `nude` nude and compared them with the entire vocabulary (over 30,000 words) from the T5 default tokenizer. Using cosine similarity, we identified the top 3 closest synonyms based on semantic embeddings. However, the results were far from rational, indicating that T5's word-level embeddings are unreliable for this task and cannot effectively evaluate semantic similarity.

Another critical issue stems from the size of T5 embeddings. With a shape of $\texttt{max\_sequence\_length}(256) \times 4096$, T5 embeddings are approximately **18 times larger** than CLIP embeddings, which have a shape of 77×768. This substantial size difference makes the adaptive selection of adversarial prompts from the vocabulary computationally intensive and time-consuming for each iteration. Noted that the goal of selecting adversarial prompts is to optimize the model, ensuring it robustly erases the target concept.

As a result, implementing semantic feature-based adversarial prompt selection in Flux incurs exceptionally high computational costs, posing a significant practical challenge for real-world applications.

Cross attention exploration: Inspired by (Hertz et al., 2022; Xie et al., 2023), we formulated a hypothesis: Does Flux exhibit a similar pattern where explicit cross-attentions exist between the given text prompt and intermediate attention maps within the network? As detailed in **Appendix A**, Flux lacks explicit cross-attention layers. Initially, this presented some challenges. However, through an in-depth examination of the neurons and features within Flux, we ultimately demonstrated (as shown in Fig. 2) that a linear relationship between text embeddings and attention maps also exists in Flux.

Specifically, as shown in Eq. (1), the feature correlation $\mathbf{Q}, \mathbf{K}$ is established by concatenating the textual and pixel embeddings along the last dimension:

$$\begin{aligned} \mathbf{Q} &= \texttt{concat}(\mathbf{Q}_{text}, \mathbf{Q}_{pixel}, \texttt{dim} = -1), \\ \mathbf{K} &= \texttt{concat}(\mathbf{K}_{text}, \mathbf{K}_{pixel}, \texttt{dim} = -1), \\ \mathbf{W}_{\mathbf{attn}} &= \texttt{Softmax}(\mathbf{Q} \times \mathbf{K}). \end{aligned} \tag{1}$$

According to our experiments, we find that the relationship between text and image is inherently forged within the confines of $\mathbf{W}_{\mathbf{attn}}$. By pinpointing the token index of the target word (want to erase) nestled within the prompt, we are capable of delineating prompt-specific characteristics. This is achieved by nullifying the pertinent column of $\mathbf{W}_{\mathbf{attn}}$, thereby elucidating the underlying features with precision.

So far, Not so good: As shown in Fig. 3, removing a target concept seems straightforward at first: by locating the token index of the keyword in the prompt, we can delete the corresponding index column in $\mathbf{W}_{\mathbf{attn}} \in [24, 1280, 1280]$, where $1280 = \texttt{max\_sequence\_length} + \texttt{head\_dim}$ and $24 = \texttt{attn\_heads}$ (generating image resolution of 512×512). However, our experiments reveal that this technique is ineffective against one of the rudimentary prompt attack strategies: obfuscating keywords—either by altering the input prompt with nonsensical prefixes or suffixes (**soccer** → **soccerrs**) or by introducing misspellings (**Nike** → **Nikke**). In such cases, the erasure of the attention map proves futile, making it easy to circumvent this method and still successfully generate the target concept. (For more details, please refer to **Appendix B**.)

Figure 3. **Attention map erasure** can be achieved by setting $\mathbf{W_{attn}}[:,:,idx_i] = 0, \forall i = (start, ..., end)$, where $start, end$ can be automatically localized given keyword *e.g.* "**soccer**" from input prompt "A child is kicking soccer". But this method is not generalizable when prompt is slightly modified.

4. Method

4.1. Overview

Building on our earlier analysis, we identified a critical limitation: deterministic attention map erasure is susceptible to conventional black-box attacks, making it unsuitable for our requirements. To address this issue, we shifted our focus to a learning-based approach. This method is designed to minimize the impact on the generation quality of irrelevant concepts while effectively erasing the target concept, ensuring a more robust and reliable solution.

We address this delicate balance between removal and preservation through a bi-level optimization strategy: the lower level is designed to enhance robust concept erasure, while the upper level ensures the maintenance of irrelevant concepts. This dual-objective methodology lies at the heart of our EraseAnything.

4.2. Bi-Level Finetuning Framework

LOWER-LEVEL: **Concept Erasure**

In the lower-level optimization phase, we refine the finetunable parameters of Flux through LoRA on the unlearned dataset D_{un}, which is comprised of concepts that we want to make Flux erased or unlearned.

ESD emerges as the relatively superior performer with higher negative guidance (Gandikota et al., 2023). The first sub-loss function employed in the lower-level optimization henceforth is formulated as Eq. (2):

$$\mathcal{L}_{esd} = \mathbb{E}\Big[v_{\theta_o+\Delta\theta}(x_t, c_{un}, t) - \eta \left\| v_{\theta_o}(x_t, c_{un}, t) - v_{\theta_o}(x_t, \emptyset, t) \right\|_2^2\Big], \tag{2}$$

where η represents the negative guidance factor, which significantly influences the degree of concept erasure. θ_o denote the parameters of the original Flux model and $\Delta\theta$ is the learnable LoRA weights for concept erasure. x_t is the denoised latent code at timestep t started with random noise at x_T (T is the total timesteps in the denoising process), $v(x_t, \emptyset, t)$ is the unconditional generation initiated with empty input prompt (*a.k.a* $\emptyset$ = `null text`), while $c_{un} \in D_{un}$ identifies the specific concept intended for erasure, for instance, *nudity*. Additionally, the term v is represent the *velocity* of the Flow matching process, which is the core part of Flux's scheduling mechanism and thus conceptually equivalent with the $v-prediction$ (Salimans & Ho, 2022) in DMs.

Furthermore, building on the insights gleaned from the cross-attention explored in Sec. 3, we strive to diminish the model's activations of the erased (unlearned) concepts by attenuating the attention weight allocated to keywords within the entire input prompt: $F^{un}_{idx} = \mathbf{W_{attn}}[:,:,idx]$.

$$\mathcal{L}_{attn} = \sum_{idx=start}^{end} F^{un}_{idx}. \tag{3}$$

Initially, we encountered suboptimal results because the fixed index positions of sensitive words, which we aimed to eliminate, could lead to overfitting. To counteract this, we scrambled the order of the sentences, thereby making the index positions dynamic. This method is reasonable because Flux can produce the similar content with a sentence that has been randomly shuffled. For more details, please refer to **Appendix B**.

UPPER-LEVEL: **Irrelevant Concept Preservation**

In the upper level, it serves for preserving concepts, which is fairly easy to understand: given the prompt c 'a nude girl...', our objective is to eliminate the word c_{un} 'nude' inside of prompt while ensuring the model can still generate an image of a unrelated concept c_{ir} normally, *e.g.* girl. To achieve this, we generate 6-10 images I_f from a fixed c and random seed (starting point of trajectory, same as DMs) that includes the concept to be removed (nude) and irrelevant concepts (girl), then train a LoRA (Low-Rank Adaptation) to induce shifts in the image generation process.

$$\mathcal{L}_{lora} = \mathbb{E}\left[\left\| v - v_{\theta+\Delta\theta}(u_t, c, t) \right\|_2^2\right], \tag{4}$$

where $v = x_T - u_{pix}$, where $x_T \sim \mathcal{N}(0, I)$ and u_{pix} is the VAE (Kingma, 2013) encoded latent code of image sampled from I_f and $u_t = (1-t)u_{pix} + tx_T$ is the noised u_{pix} at timestep t.

Apparently, for a broader range of irrelevant concepts, such as the abstract artistic styles and relationships mentioned earlier, this simple training recipe is insufficient to perserve the broader range of concepts that are not involved in the sentence. Considering the analysis in Sec. 3, explicitly incorporating a collection of images and corresponding prompt lists for irrelevant concepts is cumbersome, and T5 feature is not precise enough to measure word-level similarity.

To address this, we propose a contrastive learning approach based on the attention map of keywords. This method does not require providing a set of images corresponding to irrelevant concepts. Instead, it leverages the powerful comprehension abilities of LLMs, to heuristically generate D_{ir} that are irrelevant to the targeted concept for erasure.

First, we construct a simple AI Agent that build upon on GPT-4o to sample $c_{ir} \in D_{ir}$. For efficiency reason, we then use NLTK generating the synonym of the concept that aimed to be erased, *i.e.* the synonym of "nude" could be "nake". Specifically, we choose `K` (default is 3) irrelevant concepts. Moving forward, we fix the sampling starting latent, i.e., x_T as a constant value, and then substitute "nude" with "nake", $c_{ir}^i, i = \{1, 2, 3\}$ into c, proceeding with the denoising process independently (For more details about the c_{ir} sampling, please refer to the **Appendix C**).

As shown in Fig. 2, we choose the attention map at higher timesteps for accurate concept-related activations. Here we get the central concept's attention feature F^{un} alongside with synonym feature F^{syn} and irrelevant concept set $F^{ir} = \{F^{k_1}, ..., F^{k_K}\}$.

Drawing inspiration from the works in (Oord et al., 2018; He et al., 2020; Huang et al., 2024), we have tailored the contrastive loss to function in the opposite direction, *a.k.a*: **R**everse **S**elf **C**ontrastive loss (**RSC**): our training goal is to align the central feature F^{un} with the dynamically shifting F^{ir}, while simultaneously pushing them apart from the synonym feature F^{syn}. The strategy here is to deviate from the conventional self-contrastive learning approach, which would typically aim to make F^{un} more akin to F^{syn}, thereby enhancing the model's sensitivity to the term slated for removal. By inverting this approach, we aim to steer the network towards gradually discarding the concept of "nude" during learning, effectively obfuscating it within an array of irrelevant concepts.

$$\mathcal{L}_{rsc} = \log\left(\frac{\sum_{i=0}^{K}\exp\left(\frac{F^{un}\cdot F^{k_i}}{\tau}\right)}{\exp\left(\frac{F^{un}\cdot F^{syn}}{\tau}\right)}\right). \tag{5}$$

Algorithm 1 BO formulation in EraseAnything

Input: unlearned concept dataset and irrelevant dataset D_{un} and D_{ir}, learning rates $\alpha_{low}, \alpha_{up}$, total iteration steps M.
for $iteration = 1$ **to** M **do**
  for c_{un} sampled from D_{un} **do**
    PREPARATION
    ❶ Construct a meaningful sentence c involve c_{un}.
    ❷ Shuffle c to avoid overfitting.
    ❸ Find tokenized index $idx_{start} : idx_{end}$ of c_{un} from c.
    LOWER LEVEL: c_{un} ERASURE
    ❹ Update LoRA $\Delta\theta$ with Eq. (2)+Eq. (3) under α_{low}.
    UPPER LEVEL: c_{ir} PRESERVING
    ❺ Retrieve c_{ir}, c_{syn} *w.r.t* to c_{un} and replace them into c separately to have $F^{ir,syn}$.
    ❻ Update LoRA $\Delta\theta$ with Eq. (4)+Eq. (5) under α_{up}.
  end for
end for

As depicted in Eq. (5) (detailed derivations are provided in **Appendix D**), τ is the temperature hyperparameter that governs the model's capacity to differentiate between irrelevant concepts. A high τ causes the contrastive loss to treat all irrelevant concepts with equal importance, potentially resulting in a lack of focus in the model's learning process. Conversely, a low τ may cause the model to concentrate excessively on especially challenging irrelevant concepts, which could be mistaken for potential synonym sample. Based on empirical testing, we have determined that setting $\tau = 0.07$ is optimal for our model's performance.

Bi-Level Optimization: As shown in Eq. (5), the last loss term defined in our method is finalized. Integrating the aforementioned two optimization problems, we have a bi-level optimization illustrated in Eq. (6). (please check Alg. 1 for more details.)

$$\begin{aligned} &\min \mathcal{L}_{lora+rsc}(\Delta^*\theta; D_{ir}) \\ \textit{s.t.} \quad &\Delta^*\theta = \min \mathcal{L}_{esd+attn}(\Delta\theta; D_{un}) \end{aligned} \tag{6}$$

5. Experiments

Here, we conduct a comprehensive evaluation of EraseAnything, benchmarking it on various tasks, ranging from concrete to abstract: *e.g.*, soccer, architecture, car to artistic style, relationships and *etc*.

5.1. Implementation Details

We have opted for the Flux.1 [dev] model with publicly accessible network architecture and model weights, a distilled version of Flux.1 [pro] that retains high quality and

Table 2. Assessment of Nudity Removal: (Left) Quantity of explicit content detected using the NudeNet detector on the I2P benchmark. (Right) Comparison of FID and CLIP on MS-COCO. The performance of the original Flux [dev] is presented for reference.

METHOD	DETECTED NUDITY (QUANTITY)				MS-COCO 10K	
	COMMON	FEMALE	MALE	TOTAL↓	FID↓	CLIP↑
CA (MODEL-BASED) (KUMARI ET AL., 2023)	253	65	26	344	22.66	29.05
CA (NOISE-BASED) (KUMARI ET AL., 2023)	290	72	28	390	23.07	28.73
ESD (GANDIKOTA ET AL., 2023)	329	145	32	506	23.08	28.44
UCE (GANDIKOTA ET AL., 2024)	122	39	12	173	30.71	24.56
MACE (LU ET AL., 2024A)	173	55	28	256	24.15	29.52
EAP (BUI ET AL., 2024)	287	86	13	386	22.30	29.86
META-UNLEARNING (GAO ET AL., 2024)	355	140	26	521	22.69	29.91
OURS	129	48	22	199	21.75	30.24
FLUX.1 [DEV]	406	161	38	605	21.32	30.87

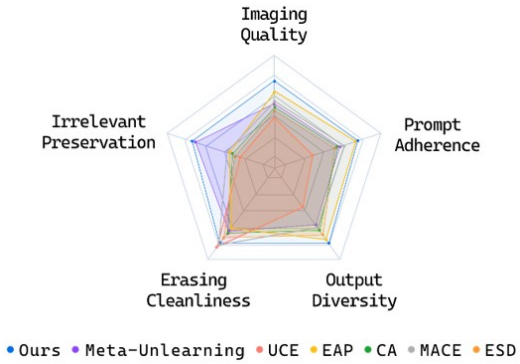

Figure 4. **User Study**. We have created an interface (see **Appendix E** for details) that shows the users with AIGC contents under various methods that transplanted to Flux. With a scoring system where 1 (worst) and 5 (best).

strong prompt adherence. Our codebase utilizes widely adopted diffusers (von Platen et al., 2022), a popular choice among developers and researchers for DMs. Unless otherwise specified, our experiments employ the flow-matching Euler sampler with 28 steps and AdamW (Loshchilov et al., 2017) optimizer for 1,000 steps, with a learning rate $\alpha_{low} = 0.001, \alpha_{up} = 0.0005$ and an erasing guidance factor $\eta = 1$ under all conditions.

In terms of concept construction, we harness the power of NLTK (Bird et al., 2009) to generate synonym concepts, and GPT-4o in the extraction of irrelevant concepts. Our fine-tuning process focuses on the text-related parameters `add_q_proj` and `add_k_proj` (subsets of $\mathbf{Q}$ and $\mathbf{K}$) within the dual stream blocks.

5.2. Results

Nudity Erasure. To assess the effectiveness and versatility of our approach, we begin by applying it to the classical task of nudity erasure. Specifically, we used our concept-erased model to generate images from a comprehensive set of 4,703 prompts extracted from the Inappropriate Image Prompt (I2P) dataset (Schramowski et al., 2023). For the identification of explicit content within these images, we deploy NudeNet (Bedapudi, 2019), using a detection threshold of **0.6**. Furthermore, to evaluate the specificity of our method in regular content, we randomly select 10,000 captions from the MS-COCO captioning dataset (validation) (Lin et al., 2014). Finally, we generate images from these captions and assess the results using both the Fréchet Inception Distance (FID) and CLIP scores.

Table 2 presents our results in comparison with the current state-of-the-art algorithms. It is evident that our method generates the second-lowest amount of explicit content when conditioned on 4,703 prompts, only outperformed by the UCE. Yet, it stands out with remarkable FID and CLIP scores, suggesting that our approach exerts a minimal negative influence on the original model's ability to generate regular content. In contrast, the UCE, while leading in explicit content reduction, shows a sharp decline in efficacy according to these metrics.

Miscellaneousness Erasure. In this section, we evaluate our method on 3 conceptual categories: *Entity*, *Abstraction*, and *Relationship*. Here, we choose 10 concept for each category (Please check Appendix C for the full list of concepts) and adopt the measuring metrics described in Table 3. As shown in Fig. 5 and Fig. 7, our method can effectively remove a variety of concepts (including multiple-concepts!) while maintaining minor disturbance compared to CA, which substantiates the claim: **EraseAnything is truly an "Erase Anything" solution**.

Table 3 reveals that our method outperforms the traditional CA in terms of erasure efficacy, the retention of unrelated concepts, and the robustness against synonym substitution. This underscores the ability of our method to not only grasp the targeted concepts for erasure but also to discern those that are semantically adjacent, all while exerting an imperceptible negative influence on the model's intrinsic capabilities. For a comprehensive evaluation of our model's robustness, kindly refer to **Appendix B**.

Figure 5. **Single-concept erasure**. We test our model across three levels of granularity—Entity, Abstraction, and Relationship—to assess its effectiveness. Furthermore, we have incorporated the versatile CA (Kumari et al., 2023) [model] to enhance the visual contrast for a clearer comparison.

Erasing "Anything". To further demonstrate "erase anything" solution requires additional supporting evidence. As shown in Fig. 6, we show the effectiveness through diverse and challenging cases, such as erasing color (i.e., red rose, green bag) or object count (i.e., five pencils, three cats).

Adversarial Attacks. Following the AdvUnlearn (Zhang et al., 2024b), we used NudeNet (Bedapudi, 2019) with a detection threshold of **0.6** to test the Attack Success Rate (ASR) on the Ring-A-Bell-Nudity dataset[1] (comprising 285 Ring-A-Bell revised prompts focused on nudity). Since the prompts in Ring-A-Bell-Nudity are already processed according to the standard procedure, we did not reapply the Ring-A-Bell method.

Table 4 shows the results of our tests on ESD, CA, and our proposed method using this dataset (all on Flux [dev]). We also included the attack results from MU-Attack (Zhang et al., 2024c). Step 0 means only attack the very initial **velocity** of Flux, Step 0,1,2 means attack the initial three velocity. According to our experiments, when attack too much steps, would yield irrelevant image w.r.t to the prompt.

User Study. To gauge the human perception of the effectiveness of our method, we conducted a user study with five dimensions, where each focusing on a different aspect of erased model. For the first two trials: Erasing Cleanliness (prompt with c_{un} and generated images do not contain concept around c_{un}) and Irrelevant Preservation (prompt with c_{ir} can be normally generated), we utilized the same concepts categorized under Entity, Abstraction, and Relationship. For each concept, images were generated using the same random seed across all methods.

[1]https://github.com/chiayi-hsu/Ring-A-Bell

Table 3. Evaluation of Erasing the specific category: Entity (*e.g.* soccer), Abstraction (*e.g.* artistic style) and Relationship (*e.g.* kiss) are presented. CLIP classification accuracies are reported for each erased category in three sets: the erased category itself (Acc_e, efficacy), the remaining unaffected categories (Acc_{ir}, specificity) and synonyms of the erased class (Acc_g, generality). All presented values are denoted in percentage (%).

METHOD	$\text{ACC}_e \downarrow$	$\text{ACC}_{ir} \uparrow$	$\text{ACC}_g \downarrow$
CA (ENTITY)	14.8	89.2	27.3
CA (ABSTRACTION)	25.2	88.3	29.6
CA (RELATIONSHIP)	22.7	88.6	23.1
OURS (ENTITY)	12.5	91.7	18.6
OURS (ABSTRACTION)	21.1	90.5	24.7
OURS (RELATIONSHIP)	18.4	90.2	19.3

Our study involved 20 non-artist participants, each providing an average of 200 responses. Fig. 4 shows that our method exhibited a comprehensive performance, achieving outstanding results across all 5 aspects, thus making EraseAnything a good all-round player in concept erasure.

5.3. Ablation study

To assess our loss design, we conducted an ablation study on the task of celebrity image erasure. We chose a subset from the CelebA (Liu et al., 2018), omitting those that Flux [dev] couldn't accurately reconstruct. This resulted in a dataset of 100 celebrities, split into two groups: 50 for erasure and 50 for retention. Unlike MACE's massive concept erasure, EraseAnything is trained on individual celebrities.

Different variations and their results (evaluated by averaging metrics) are presented in Table 5. $\mathcal{L}_{esd}$ itself fall short of the complete erasure of target concept, resulting in a not so low ACC_e. With the addition of $\mathcal{L}_{attn}$, ACC_e has fallen dramatically but the retention of irrelevant concepts was fail *w.r.t* ACC_{ir}. Incorporating the loss term $\mathcal{L}_{rsc}$, we introduce a approach that may lead to achieving high ACC_{ir} values. By organically combining all these loss terms, we achieve a comprehensive model that consistently demonstrates the lowest ACC_e and the highest ACC_{ir} compared to previous configurations.

Others. Due to the page limits, we put remaining experimental details and results in **Appendix F**. This includes the visualizations under different configs and objects and more.

Table 4. Performance Metrics of Nudity Detection Methods.

Concept	Methods	Flux[dev]	ESD	CA	EraseAnything
Nudity	Original (Org)	59.65%	7.36%	3.16%	2.46%
	MU-Attack (step 0)	64.56%	11.57%	15.44%	8.77%
	MU-Attack (steps 0,1,2)	65.96%	14.74%	16.49%	11.93%

Table 5. Ablation Study on Erasing Celebrities, we ablate four loss terms used in our experiments.

CONFIG	$\text{ACC}_e \downarrow$	$\text{ACC}_{ir} \uparrow$
$\mathcal{L}_{esd} + \mathcal{L}_{attn}$	15.3	82.1
$\mathcal{L}_{esd} + \mathcal{L}_{lora}$	20.5	77.9
$\mathcal{L}_{esd} + \mathcal{L}_{rsc}$	16.1	85.6
$\mathcal{L}_{attn} + \mathcal{L}_{rsc}$	18.6	81.7
$\mathcal{L}_{attn} + \mathcal{L}_{lora} + \mathcal{L}_{rsc}$	15.8	80.2
FULL	**14.9**	**88.5**

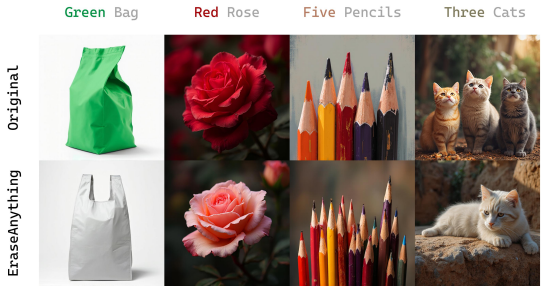

Figure 6. **Erase complex concepts**.

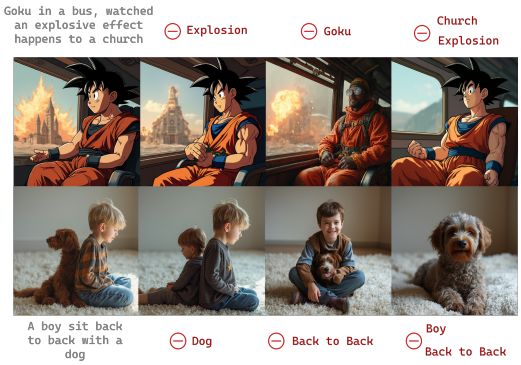

Figure 7. **Multi-concept erasure**.

6. Conclusion

In this paper, we propose **EraseAnything**, a Flux-based concept erasing method. Leveraging a bi-level optimization strategy, we strike a balance between erasing the target concept that bound to be removed while preserving the irrelevant concepts unaffected, mitigating long-lasting notorious risk of overfitting and catastrophic forgetting. Experiments across diverse tasks strongly demonstrate the effectiveness and versatility of our method.

Impact Statement

Ethical Aspects

Potential Risks of Content Generation: The method proposed in this paper enables the removal of unwanted concepts from large-scale text-to-image (T2I) diffusion models, such as inappropriate content. This helps to reduce the risk of generating content that violates copyright, privacy, or is otherwise unsuitable, thereby mitigating potential ethical controversies associated with the use of such models. Specifically, it addresses concerns regarding the generation of NSFW (Not Suitable For Work) material, which has been a significant issue highlighted in various reports.

Fairness and Bias of Models: By precisely erasing specific concepts, the model's output can be more finely controlled, avoiding the generation of unfair or biased content that may arise due to biases in the training data. This ensures that the model serves different users and application scenarios in a more equitable and neutral manner.

Societal Consequences

Content Creation and Moderation: With the increasing application of T2I models in content creation, the method presented in this paper provides content creators and moderators with an effective tool to control the subject matter and style of generated content more accurately. This improves the quality and safety of content creation and also alleviates the workload of content moderation, contributing to a healthier and more positive online content ecosystem.

Model Interpretability and Trust: Through the research and implementation of concept erasure techniques, a better understanding and interpretation of the working principles of T2I models can be achieved. This understanding of how models learn and generate specific content enhances users' trust in the models, facilitating their wider adoption and application in various fields such as education, healthcare, and artistic creation, thereby bringing more innovation and value to society.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant 62121002, U2336206, 62372423, 62441617 and 62102386. It was also supported by the Beijing Natural Science Foundation under Grant No. 4254100, the Fundamental Research Funds for the Central Universities under Grant No. KG16336301, the China Postdoctoral Science Foundation under Grant No. 2024M764093, and by Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority.

First author and his PHD advisors are with School of Cyber Science and Technology, University of Science and Technology of China and Anhui Province Key Laboratory of Digital Security, Hefei 230026, China.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bedapudi, P. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019.

Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Bui, A., Vuong, L., Doan, K., Le, T., Montague, P., Abraham, T., and Phung, D. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024.

Colson, B., Marcotte, P., and Savard, G. An overview of bilevel optimization. *Annals of operations research*, 153: 235–256, 2007.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

Gao, H., Pang, T., Du, C., Hu, T., Deng, Z., and Lin, M. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*, 2024.

Hao, Z., Ying, C., Su, H., Zhu, J., Song, J., and Cheng, Z. Bi-level physics-informed neural networks for pde constrained optimization using broyden's hypergradients. *arXiv preprint arXiv:2209.07075*, 2022.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, C.-P., Chang, K.-P., Tsai, C.-T., Lai, Y.-H., and Wang, Y.-C. F. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023.

Huang, Z., Wu, T., Jiang, Y., Chan, K. C., and Liu, Z. ReVersion: Diffusion-based relation inversion from images. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Li, L., Lu, S., Ren, Y., and Kong, A. W.-K. Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. *arXiv preprint arXiv:2504.12782*, 2025.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Liu, Y., An, J., Zhang, W., Li, M., Wu, D., Gu, J., Lin, Z., and Wang, W. Realera: Semantic-level concept erasure via neighbor-concept mining. *arXiv preprint arXiv:2410.09140*, 2024.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020.

Loshchilov, I., Hutter, F., et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

Lu, S., Liu, Y., and Kong, A. W.-K. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2294–2305, 2023.

Lu, S., Wang, Z., Li, L., Liu, Y., and Kong, A. W.-K. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024a.

Lu, S., Zhou, Z., Lu, J., Zhu, Y., and Kong, A. W.-K. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024b.

Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., and Ding, G. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rando, J., Paleka, D., Lindner, D., Heim, L., and Tramèr, F. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

Shen, Q., Wang, Y., Yang, Z., Li, X., Wang, H., Zhang, Y., Scarlett, J., Zhu, Z., and Kawaguchi, K. Memory-efficient gradient unrolling for large-scale bi-level optimization. *arXiv preprint arXiv:2406.14095*, 2024.

Sinha, A., Malo, P., and Deb, K. A review on bilevel optimization: From classical to evolutionary approaches and applications. *IEEE transactions on evolutionary computation*, 22(2):276–295, 2017.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., and Shou, M. Z. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

Zhang, L., Liang, Y., and Xie, P. Blo-sam: Bi-level optimization based overfitting-preventing finetuning of sam. *arXiv preprint arXiv:2402.16338*, 2024a.

Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., and Liu, S. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *arXiv preprint arXiv:2405.15234*, 2024b.

Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., and Liu, S. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024c.

A. Flux Architecture

In our research, we have chosen Flux [dev] as our baseline model due to its reputation as the most performant within the open-source Flux series [2]. As highlighted in Section 3, Flux's architecture significantly diverges from that of SD v1.5, which has been the predominant baseline for contemporary concept erasure techniques.

As shown in Figure 8 and Figure 9, we have dissected the architecture of Flux ([schnell] and [dev] shared the same architecture). We discovered that, unlike in SD, Flux does not incorporate an explicit cross-attention module. Nonetheless, we have observed that the dual stream block's approach to concatenating text and image features can emulate the cross-attention effects of SD. Specifically, this mechanism enables the identification of a word's heatmap within the attention map based on the token's position in the text, which can be seen in Figure 10. Furthermore, we have found that by pruning this heatmap, we can effectively inhibit the generation of specific content, a finding that serves as a pivotal foundation in our paper.

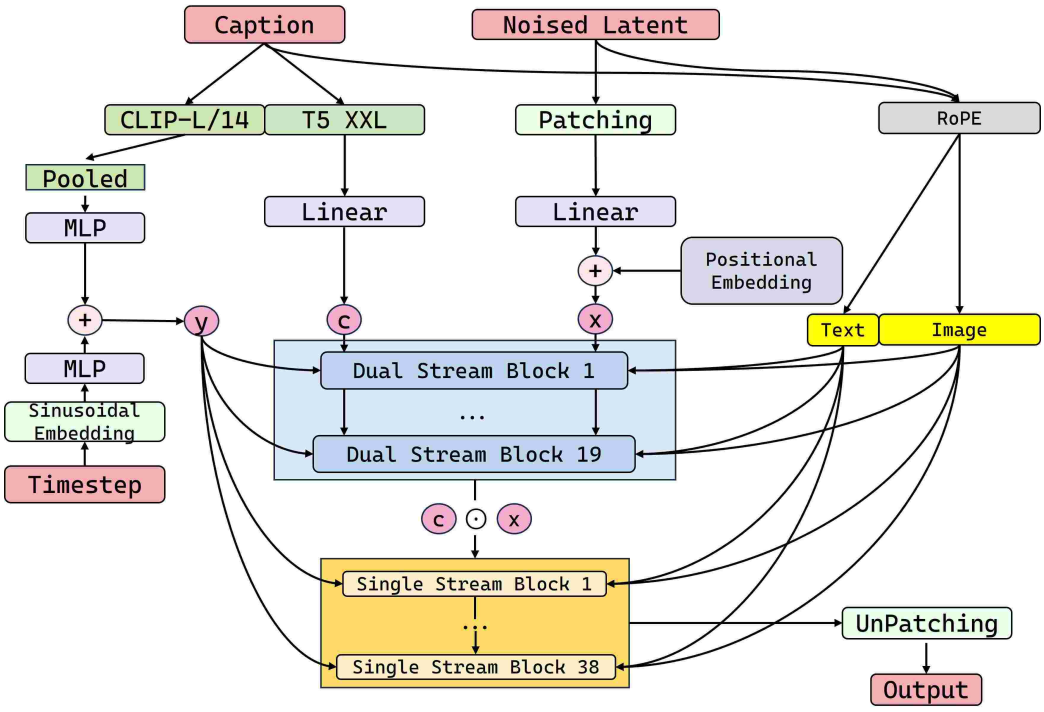

Figure 8. **Model architecture of Flux [dev]**. Flux [dev] use frozen CLIP-L 14 and T5-XXL as text encoders for conditioned caption feature extraction. The coarsed CLIP embedding concatenated with timestep embedding y are used to modulation mechanism. The fine-grained T5 c concatenated with image latents x are input to a stacked of double stream blocks and single stream blocks to predict output in the VAE encoded latent space. Concatenation is indicated by $\odot$.

Building upon this finding, our optimization efforts are now focused on the dual stream block, as illustrated in Figure 9). Our experimental results indicate that the parameters `add_v_proj` and `to_v` are highly numerically sensitive, rendering them less than ideal for optimization purposes. Consequently, we have shifted our focus to optimizing `add_q(k)_proj` and `to_q(k)` instead. This strategic adjustment is expected to yield more robust and stable improvements in the model's performance.

For a fair comparison, we have adapted traditional methods such as ESD, UCE, and MACE, which typically optimize the $\mathbf{Q}, \mathbf{V}$, to instead optimize the $\mathbf{Q}, \mathbf{K}$ inside of Dual Transformer Block. This modification ensures that our comparative analysis is conducted under a consistent and relevant framework.

[2] https://blackforestlabs.ai/announcing-black-forest-labs/

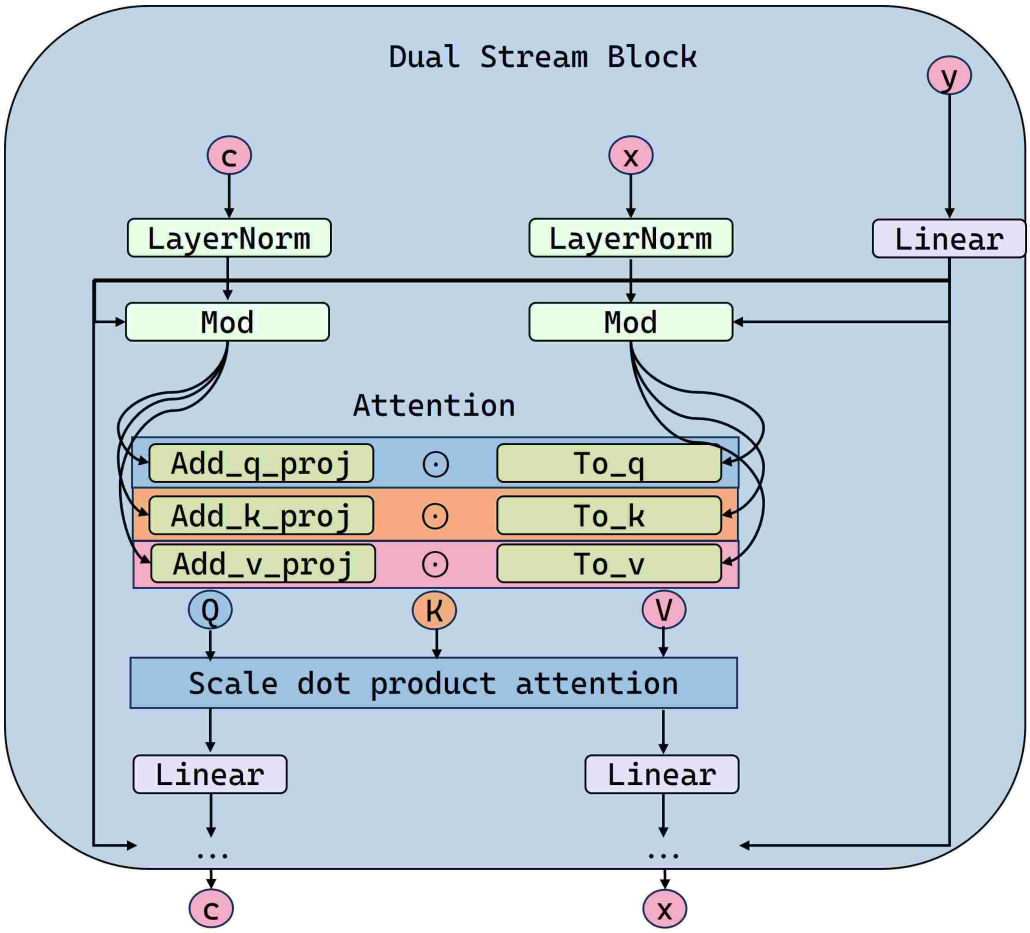

Figure 9. **Dual stream block**. In Flux, the semantic correlation is established in the dual stream block, which established an implicit relationshio between text and image. Noteworthy thing is that the explicit cross attention module that prevails among SD v1.5 is not existed in Flux.

Figure 10. **Attention map extraction**. The correlation between specific words and their corresponding heatmaps can be discerned within the matrix $\mathbf{W}_{\mathbf{attn}}$, particularly within the columns (**white** bar adorned with a **blue** dotted line) associated with text.

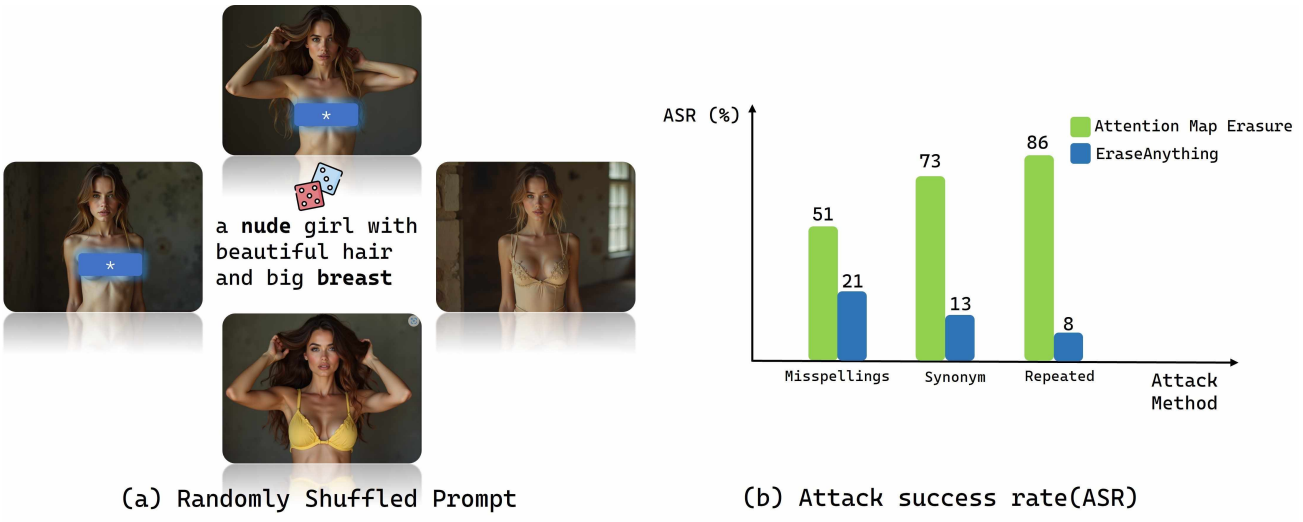

Figure 11. **Order Insensitive & Black box attack**. (a) The sequence of the prompt has minimal impact on the synthesized image. (b) Our learning-based method can maintain robustness against conventional black box attacks, whereas attention map erasure is ineffective.

B. Pattern of prompt & Black box attack

To address the issue of overfitting, we aim to make the token index dynamic. Initially, we must validate a hypothesis: "**Randomly shuffling the prompt should not impact the generation results of Flux**".

The basic prompt in our case is: "`a nude girl with beautiful hair and big breast`". To demonstrate Flux's generalizability, we randomly shuffled this prompt at the word level: *e.g.* "`girl with beautiful and big nude a hair breast`". To ensure fairness, we fed these randomly shuffled prompts into a popular online service, Fal.ai [3]. Fal.ai is known for providing off-the-shelf Text2Image APIs in an easily accessible manner, making it popular among users who wish to quickly test their ideas and create prototypes. We chose Fal.ai due to its swift image generation capabilities and the tamper-proof nature of its model weights.

As depicted in Figure 11 (a), despite the alteration of word order within the prompt, the central attributes of the prompt remained robust: "`beautiful`; `girl`; `nude`; `hair`; `breast`" (even though the generated results oscillated between sensitive and regular content). Therefore, this experiment sufficiently demonstrated a key characteristic of Flux [dev]: **Flux [dev] is not sensitive to the word order in the input prompt**.

This serves as a compelling demonstration that we can effectively employ data augmentation by utilizing this property. It justifies the practice of shuffling the prompt at each iteration during training, enhancing the robustness of our model.

Furthermore, we have curated a set of 100 prompts that include recognizable objects or styles, spanning from soccer, celebrities, to cartoons and art. Our goal here is to verify that the simple attention map erasure technique, as discussed in the context of cross-attention in Section 3), can be easily circumvented through rudimentary black-box prompt attacks.

As illustrated in Figure 11 (b), the attention map erasure technique struggles to effectively handle misspellings and synonyms, as the token index for the target concept word differs from those of its misspellings and synonyms. Regarding the scenario where the target concept word is repeated (*i.e.*, it appears at least twice in the prompt), we have observed that the complete deletion of attention maps associated with the corresponding indices does not prevent the re-generation of the target concepts. As shown in Figure 12, the attempted deletion of "**New Balance**" and "**Dr.Martens**" does not yield the expected outcome.

This finding underscores the complexity of the task and suggests that a more sophisticated approach is needed to ensure that the target concepts are not regenerated in the output, regardless of their frequency in the input prompt. The current method of attention map erasure does not suffice, and thus, there is a clear need for a more nuanced learning-based erasure technique that can distinguish and eliminate the influence of repeated target concepts effectively. As demonstrated in Figure 11 (b), our method can effectively counter these black-box attack methods and significantly lower the **attack success rate**

[3] https://fal.ai/

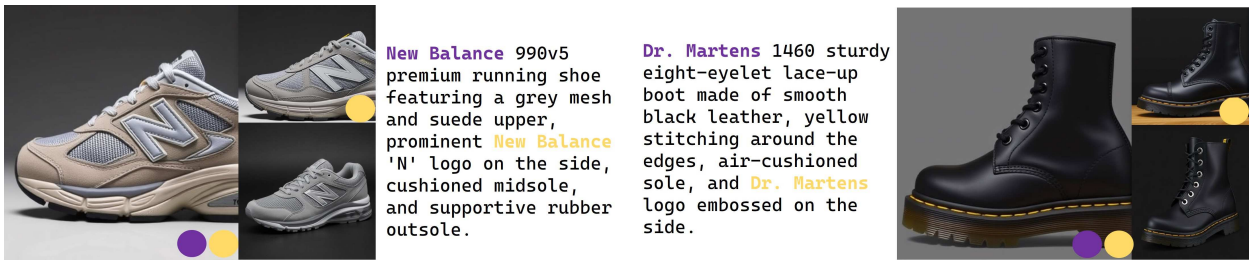

Figure 12. **Repeated (target concept occurs more than twice in the input prompt)**. It is apparent that the direct attention map erasure proves ineffective in addressing the re-generation problem of the target concept within the prompts. As illustrated in the figure, the first token index is denoted by purple, and the second token index is denoted by gold. We discovered that even after zeroing out all concept-related token indices in the attention map, the resulting image still includes the concept that was intended to be erased.

Table 6. AI Agent template in generating c_{ir} (c_{un} = "nude").

Role	**Content**
System	'*You are a helpful assistant and a well-established language expert*'
User	Hello, please return K (K=3) English words that you think with Human intuition are **no_relation/far/mid** in the semantic space from the English word: c_{un}, and only reply the result with JSON format is as follows: {"**no_relation**": [(word1, similarity_score1), ...], "**far**": [(word1, similarity_score1), ...], "**mid**": [(word1, similarity_score1), ...]}
Response	{"**no_relation**": [("cloud", 0.1), ("tree", 0.2), ("carpet", 0.1)], "**far**": [("hot", 0.3), ("color", 0.4), ("wet", 0.3)], "**mid**": [("image", 0.5), ("figure", 0.6), ("portrait", 0.5)]}

(ASR) below the acceptable level.

C. Prompt-related supplementary material

C.1. A heuristic c_{ir} sampling method

Identifying the concept c_{ir} that is unrelated to the target concept in the semantic feature space is not as straightforward as it may seem. General text feature encoders like T5 are typically trained on large-scale corpus data. The repeated occurrence of two seemingly unrelated concepts in the same training corpus might lead to a certain degree of correlation in the semantic feature dimension, causing the mapping position relationship of different text tokens in their semantic space to deviate from human perception of text words. Therefore, the similarity between text embeddings cannot be directly used as a measure to represent the correlation between two concepts.

To address this issue, we have devised a a heuristic c_{ir} sampling method. By leveraging the cognitive ability of LLM regarding human text concepts and through heuristic prompt design, we make them return concepts that are unrelated to the word c_{un} to be erased and also require the similarity between c_{un} and c_{ir}. Since the interaction with the LLMs occurs at the natural language level, the returned similarity is only a relative reference value, but it suffices to meet our requirements for sampling c_{ir}.

As shown in Table 6, the process of c_{ir} is first through building an AI Agent with unique role and regulated output format. We initiate the process by requiring GPT-4o to return c_{ir} that they deem to be unrelated to the target concept. After got the set of candidate values. Next, we classify and rank these concepts into three distinct categories: "**no_relation**", signifying concepts that have minimal or no semantic connection; "**far**", representing those with a relatively loose semantic association; "**mid**", indicating a moderate level of relatedness.

After obtaining the initial response in Table 6, we randomly select each word from the three categories, which is in accordance with K = 3 by default as illustrated in the main paper.

Table 7. Complete list of conceptions of Entity, Abstraction, Relationship

Category	# Number	Prompt template	Conceptions
Entity	10	'A photo of [*Entity*]'	'Fruit', 'Ball', 'Car', 'Airplane', 'Tower', 'Building', 'Celebrity', 'Shoes', 'Cat', 'Dog'
Abstraction	10	'An Art in the style of [*Abstraction*]'	'Pablo Picasso', 'Salvador Dali', 'Claude Monet', 'Vincent Van Gogh', 'Rembrandt van Rijn', 'Frida Kahlo', 'Edvard Munch', 'Leonardo da Vinci', 'Explosions', 'Environmental Simulation'
Relationship	10	'A [*Relationship*] B'	'Shake Hand', 'Kiss', 'Hug', 'In', 'On', 'Back to Back', 'Jump', 'Burrow', 'Hold', 'Amidst'

C.2. Complete list of Entity, Abstraction, Relationship

For assessing the generalization of EraseAnything, we establish a conception list at three levels: from the concrete objects to the abstract artistic style and relationship, the full list used in our experiments is presented in Table 7.

D. Derivative of Reverse Self-Contrastive Loss

As one of the proven method, InfoNCE loss is widely used in self-contrastive learning to learn model parameters by contrasting the similarity between positive and negative samples:

$$\mathcal{L}_{InfoNCE} = -\log\left(\frac{\exp(\text{sim}(q,k^+))}{\sum_{i=0}^{N}\exp(\text{sim}(q,k_i))}\right) \tag{7}$$

where $\text{sim}(q,k)$ denotes the similarity between the query vector q and the key vector k, k^+ is the key vector of the positive sample, k_i represents the key vectors of negative samples, and K is the number of negative samples.

In conventional self-contrastive learning, we aim to make F^{un} more similar to F^{syn} to enhance the model's sensitivity to the term targeted for removal.

$$\mathcal{L}_{sc} = -\log\left(\frac{\exp\left(\text{sim}(F^{un}\cdot F^{syn})\right)}{\sum_{i=0}^{K}\exp\left(\text{sim}(F^{un}\cdot F^{k_i})\right)}\right) \tag{8}$$

However, in our case, we desire the model to be less sensitive to the term "nude" and its synonyms. Thus, we introduce the **Reverse Self-Contrastive Loss** through swapping the numerator and the denominator:

$$\mathcal{L}_{rsc} = \log\left(\frac{\sum_{i=0}^{K}\exp(\text{sim}(F^{un},F^{k_i}))}{\exp(\text{sim}(F^{un},F^{syn}))}\right) \tag{9}$$

Here, F^{un} is the central feature, F^{syn} is the synonym feature, and F^{k_i} are the features of other irrelevant concepts.

To refine the model further, we consider introducing a temperature parameter τ to adjust the distribution of similarity scores:

$$\text{sim}(F^{un},F^{syn}) = \frac{F^{un}\cdot F^{syn}}{\tau} \tag{10}$$

Incorporating the temperature parameter into the loss function, we obtain:

$$\mathcal{L}_{rsc} = \log \left(\frac{\sum_{i=0}^{K} \exp\left(\frac{F^{un} \cdot F^{ki}}{\tau}\right)}{\exp\left(\frac{F^{un} \cdot F^{syn}}{\tau}\right)} \right) \tag{11}$$

This derivation integrates the fundamental concepts of the InfoNCE loss function and tailors them to our specific case. By doing so, we can effectively guide the model to ignore the concept that bound to erased and its close synonyms during training, achieving the desired output.

E. User Study

Adhering to Flux's comprehensive evaluative criteria for Text-to-Image (T2I) models, we have integrated three key metrics into our user study: **Imaging Quality, Prompt Adherence, Output Diversity**. These metrics serve as the cornerstone for assessing the performance of our model. In our specific context, which focuses on the erasure of concepts to minimize their interference with the synthesis of images featuring unrelated concepts, we have introduced two additional metrics to refine our assessment framework: **Erasing Cleanliness** and **Irrelevant Preservation**.

Erasing Cleanliness evaluates the effectiveness of the concept erasure process, ensuring that the targeted concepts are thoroughly removed without leaving any residual influence on the synthesized image. Irrelevant Preservation, on the other hand, measures the model's ability to maintain the integrity and relevance of concepts that are not the focus of the erasure process, ensuring that the overall composition and context of the image are preserved within the model.

Figure 13 and Figure 14 provide a visual representation of the user study interface, which was meticulously designed to facilitate a smooth and engaging participant experience. During the study, participants were presented with a series of image sets, each containing 6 and 3 results generated by various anonymous methods. They were then prompted to score each method based on its performance across the aforementioned metrics. The collected data was subsequently compiled and visualized in a pentagonal chart, as depicted in Figure 4 of the main paper, offering a comprehensive overview of the methods' performance and highlighting the strengths and areas for improvement of each approach. This visual summary serves as a valuable tool for both researchers and practitioners, enabling a more nuanced understanding of the model's capabilities and guiding future developments in the field of image synthesis.

F. Others

Important Declaration: Owing to the rigid 10MB file size limit of ICML 2025, authors have had to compress the images to a level that significantly reduces their quality in the appendices.

F.1. Celebrity

The names of celebrities used in our ablation study are illustrated in Table 8. The noteworthy thing here is that not arbitrary celebrities can be faithfully synthesised by Flux [dev], after manually comparing the synthesized famous people with its prompt and add some comic characters, we keep 50 for each group.

Specification: We train the celebrity recognition network on top of **MobileNetV2** that pretrained on ImageNet, then add a `GlobalAveragePooling2D` and `Softmax(Dense)` at the end of the orginal output (`out_relu`) of MobileNetV2. The learning rate is a fixed 1e-4 with Adam optimizer and loss function is categorical cross-entropy.

As for dataset, we gather the data with an average of 50 pictures per celebrity, with the gross number of 5,000. Then we randomly re-sampled the dataset and divided into training set (80%) and test set (20%). The statistics are reported upon the test set (1,000), reserves one decimal fraction.

F.2. More Experimental Results

Ablation Study on Diverse Loss Configurations. As demonstrated in Figure 19, we conducted a thorough comparison of outcomes utilizing various combinations of loss functions to our methodology. It is evident that the strategic integration of $\mathcal{L}_{lora}$ significantly bolsters the visual consistency with the original character's appearance. Meanwhile, $\mathcal{L}_{rsc}$ adeptly obscures the targeted concept, directing it towards transformation into a myriad of incongruous notions. In contrast, $\mathcal{L}_{esd}$ exemplifies the quintessential concept erasure strategy.

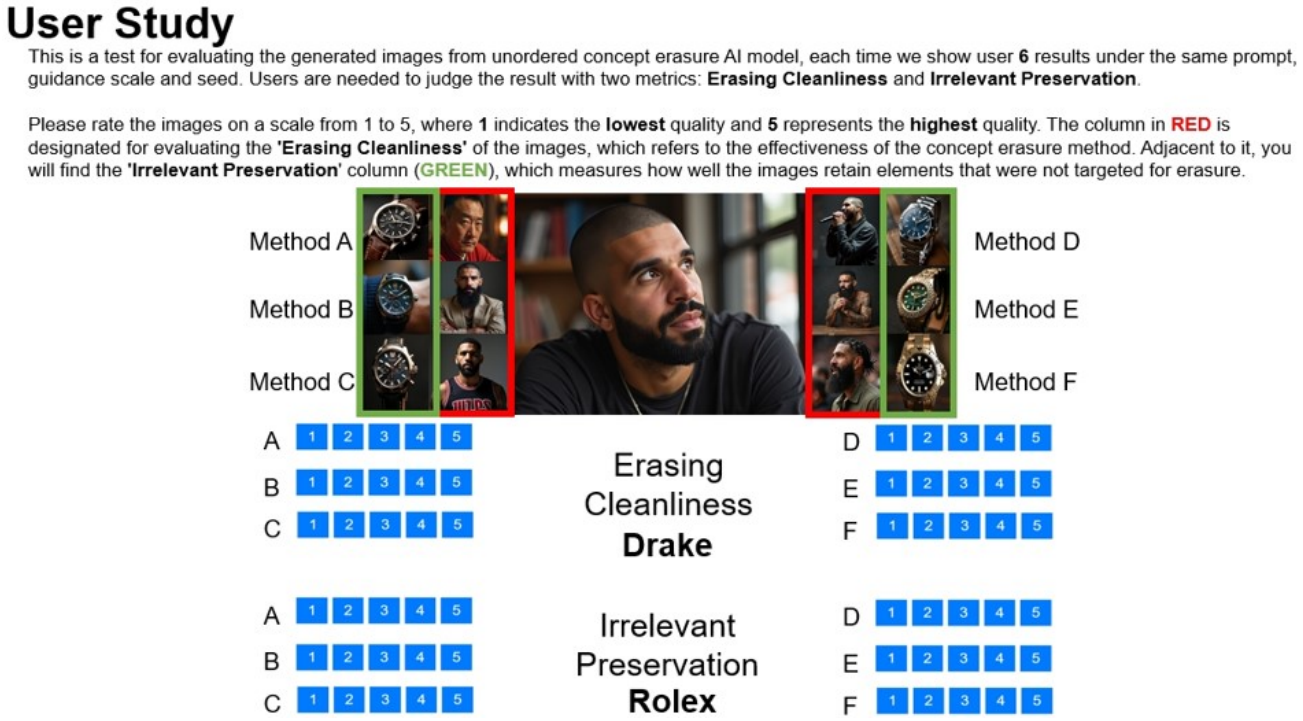

Figure 13. **User Study on Erasing Cleanliness and Irrelevant Preservation** (Screenshot).

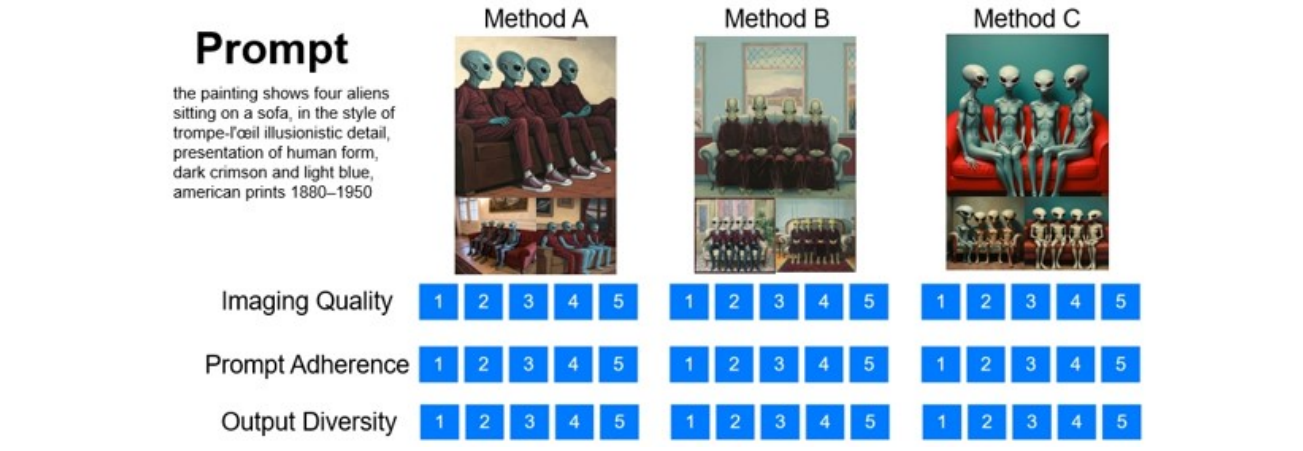

Figure 14. **User Study on Imaging Quality, Prompt Adherence and Output Diversity** (Screenshot).

Table 8. Complete list of celebrities used in ablation study.

Category	# Number	Celebrity
Erasure Group	50	'*Adele*', '*Albert Camus*', '*Angelina Jolie*', '*Arnold Schwarzenegger*', '*Audrey Hepburn*', '*Barack Obama*', '*Beyoncé*', '*Brad Pitt*', '*Bruce Lee*', '*Chris Evans*', '*Christiano Ronaldo*', '*David Beckham*', '*Dr Dre*', '*Drake*', '*Elizabeth Taylor*', '*Eminem*', '*Elon Musk*', '*Emma Watson*', '*Frida Kahlo*', '*Hugh Jackman*', '*Hillary Clinton*', '*Isaac Newton*', '*Jay-Z*', '*Justin Bieber*', '*John Lennon*', '*Keanu Reeves*', '*Leonardo Dicaprio*', '*Mariah Carey*', '*Madonna*', '*Marlon Brando*', '*Mahatma Gandhi*', '*Mark Zuckerberg*', '*Michael Jordan*', '*Muhammad Ali*', '*Nancy Pelosi*','*Neil Armstrong*', '*Nelson Mandela*', '*Oprah Winfrey*', '*Rihanna*', '*Roger Federer*', '*Robert De Niro*', '*Ryan Gosling*', '*Scarlett Johansson*', '*Stan Lee*', '*Tiger Woods*', '*Timothee Chalamet*', '*Taylor Swift*', '*Tom Hardy*', '*William Shakespeare*', '*Zac Efron*'
Retention Group	50	'*Angela Merkel*', '*Albert Einstein*', '*Al Pacino*', '*Batman*', '*Babe Ruth Jr*', '*Ben Affleck*', '*Bette Midler*', '*Benedict Cumberbatch*', '*Bruce Willis*', '*Bruno Mars*', '*Donald Trump*', '*Doraemon*', '*Denzel Washington*', '*Ed Sheeran*', '*Emmanuel Macron*', '*Elvis Presley*', '*Gal Gadot*', '*George Clooney*', '*Goku*','*Jake Gyllenhaal*', '*Johnny Depp*', '*Karl Marx*', '*Kanye West*', '*Kim Jong Un*', '*Kim Kardashian*', '*Kung Fu Panda*', '*Lionel Messi*', '*Lady Gaga*', '*Martin Luther King Jr.*', '*Matthew McConaughey*', '*Morgan Freeman*', '*Monkey D. Luffy*', '*Michael Jackson*', '*Michael Fassbender*', '*Marilyn Monroe*', '*Naruto Uzumaki*', '*Nicolas Cage*', '*Nikola Tesla*', '*Optimus Prime*', '*Robert Downey Jr.*', '*Saitama*', '*Serena Williams*', '*Snow White*', '*Superman*', '*The Hulk*', '*Tom Cruise*', '*Vladimir Putin*', '*Warren Buffett*', '*Will Smith*', '*Wonderwoman*'

Benchmarking Against State-of-the-Art (SOTA). As depicted in Figure 15, we compare EraseAnything with state-of-the-art (SOTA) methods on various concepts. It can be easily observed that **Attention Map** is sufficient to remove target concept. However, as previously analyzed in Appendix B, such methodologies are susceptible to rudimentary black-box attacks, rendering them impractical for real-world applications.

LoRA Disentanglement Analysis. To assess the potential influence of integrating fine-tuned LoRAs into the original Flux [dev], as depicted in Figure 16, it can be observed that incorporating fine-tuned LoRAs for diverse concepts, *i.e. Celebrity*: **Batman, Christiano Ronaldo, Hulk, Lebron James, Wonderwoman**. *Object*: **Alaskan Malamute, Statue of Liberty, Basketball, Skyscraper, Cat** and *Art*: **Van Gogh, Edvard Munch, Rembrandt van Rijn, Claude Monet, Salvador Dali**, does not adversely affect the original image synthesis capabilities. All above-mentioned concepts are depicted sequentially from left to right.

Exploring the Synergy of Combined Concept-Erased LoRAs. In our quest to unravel the potential of integrating concept-erased LoRAs, we delve into the intricacies of merging these elements into a cohesive single entity, denoted as $\mathbf{\Delta}\theta_{\mathbf{mul}}$. This experiment is meticulously designed to assess the capabilities of image synthesis when multiple LoRAs are unified. Specifically, we randomly sample LoRAs from Table 7 and combine them using Equation (12).

As depicted in Figure 17, the **upper** side of the blue dashed line represents $\sum_{i=0}^{N} W_i = 1, W_i = \frac{1}{N}$, indicating a linear normalized weight blending strategy. Conversely, the **lower** side of the line reveals the implications of a non-normalized sum, where $\sum_{i=0}^{N} W_i = N, W_i = 1$. Here, N represents the total number of LoRAs being combined, *e.g.* **3, 5, 10**.

$$\mathbf{\Delta}\theta_{\mathbf{mul}} = \sum_{i=0}^{N} W_i \Delta\theta_i \tag{12}$$

The process of image synthesis is significantly impacted when the cumulative weight of the combined LoRAs, denoted by $\sum_{i=0}^{N} W_i$, exceeds the normalized threshold of 1. This surpassing signals a critical juncture in the image synthesis process, potentially resulting in an overemphasis on certain concepts while inadvertently neglecting others. Such a shift could introduce a bias towards recognized concepts, possibly at the expense of exploring new or unrelated themes.

Conversely, when the aggregate weight remains within the confines of 1, the model's prowess in generating a diverse array of unrelated concepts remains largely indistinguishable from the original Flux[dev] model, underscoring the model's robustness.

Multiple Concept Erasure. Leveraging the insights gleaned from aforementioned findings, we venture to explore the hypothesis of concept erasure with greater depth:

`Q: Can EraseAnything is capable of erasing multiple concepts in the meantime?`

Resoundingly, the answer is affirmative. As depicted in Figure 18, through the linear interweaving of LoRAs representing distinct concepts under a normalized weight sum, we achieve the coveted outcome of concept erasure that harmoniously integrates with the backdrop of the environment. This capability positions EraseAnything as an exemplary contender for advanced concept erasure endeavors.

F.3. Limitations

Although "EraseAnything" has demonstrated its formidable ability to erase concepts across various domains, we have identified challenges it faces in certain situations:

Extensive Concept Erasure: When tasked with erasing multiple concepts simultaneously, such as 10 or more concepts (LoRAs), the **Normalized Sum** strategy, as depicted in Equation (12), results in a proportional decrease in the impact of each concept's erasure. Consequently, a significant and important avenue for research in this field is to explore efficient methods for combining a large number of LoRAs (more than 100) effectively.

Fine-grained Control: Another issue pertains to the inability to guarantee the strength of the erasure during fine-tuning. This is an uncharted yet intriguing area in the realm of concept erasure, which could provide us with a deeper understanding of the concept formulation. It would also enable more precise control over the erasure process, *e.g.* a slider could be provided to control the intensity during interactive concept erasure.

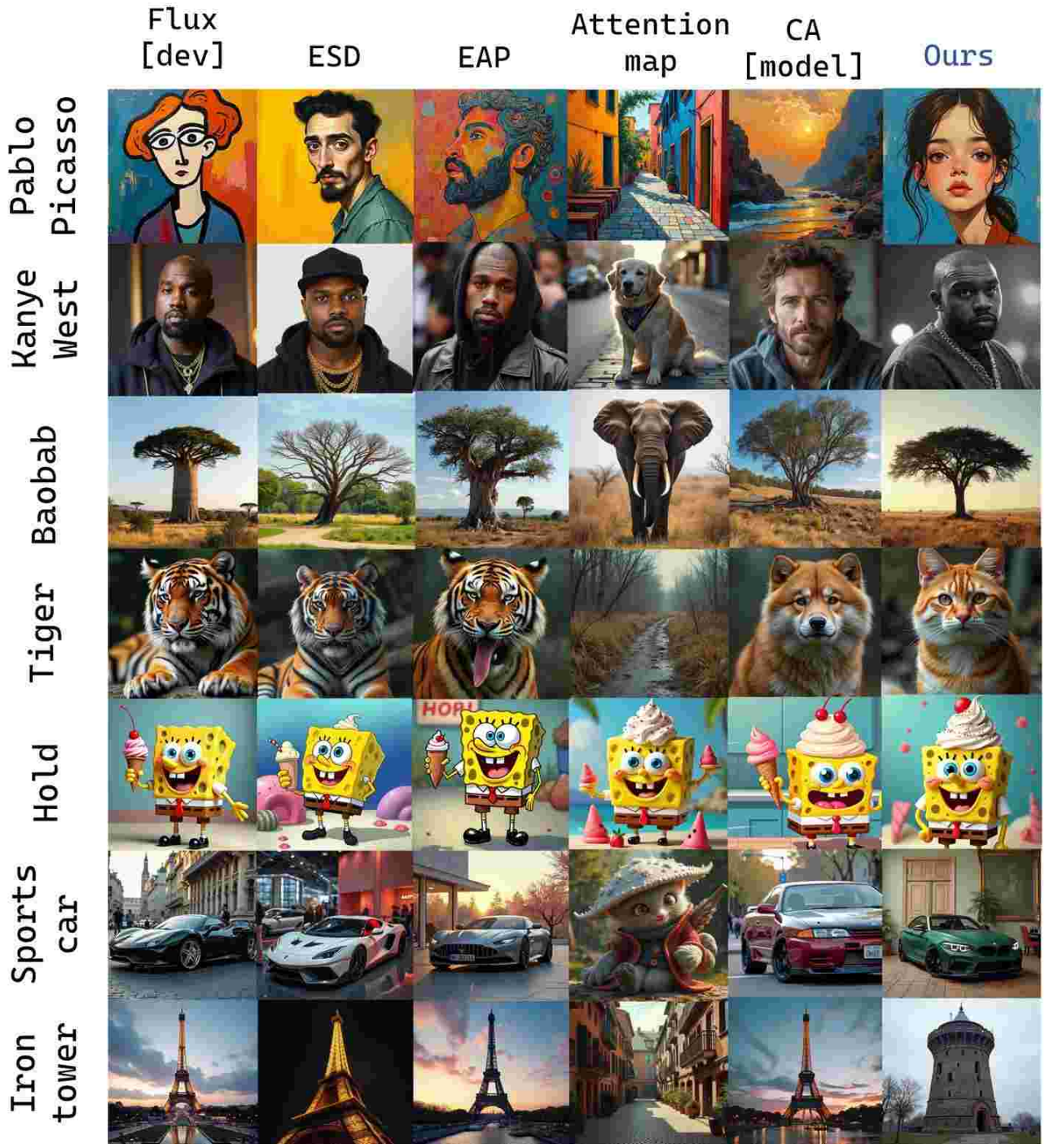

Figure 15. **Comparison with mainstream concept erasing methods**. We compared EraseAnything to other concept erasers on Flux [dev] across categories like *Art Style*, *Celebrity*, *Plant & Animal*, *Relationship*, *Car & Architecture*. The **Attention Map** (3rd column from the right) shows the simple token localization method from Section 3 that erases target concept effectively, yet its vulnerable to the minor change of tokens———*misspellings, prefixes & suffixes and repeated words*———make it difficult to widely adopt in practical applications.

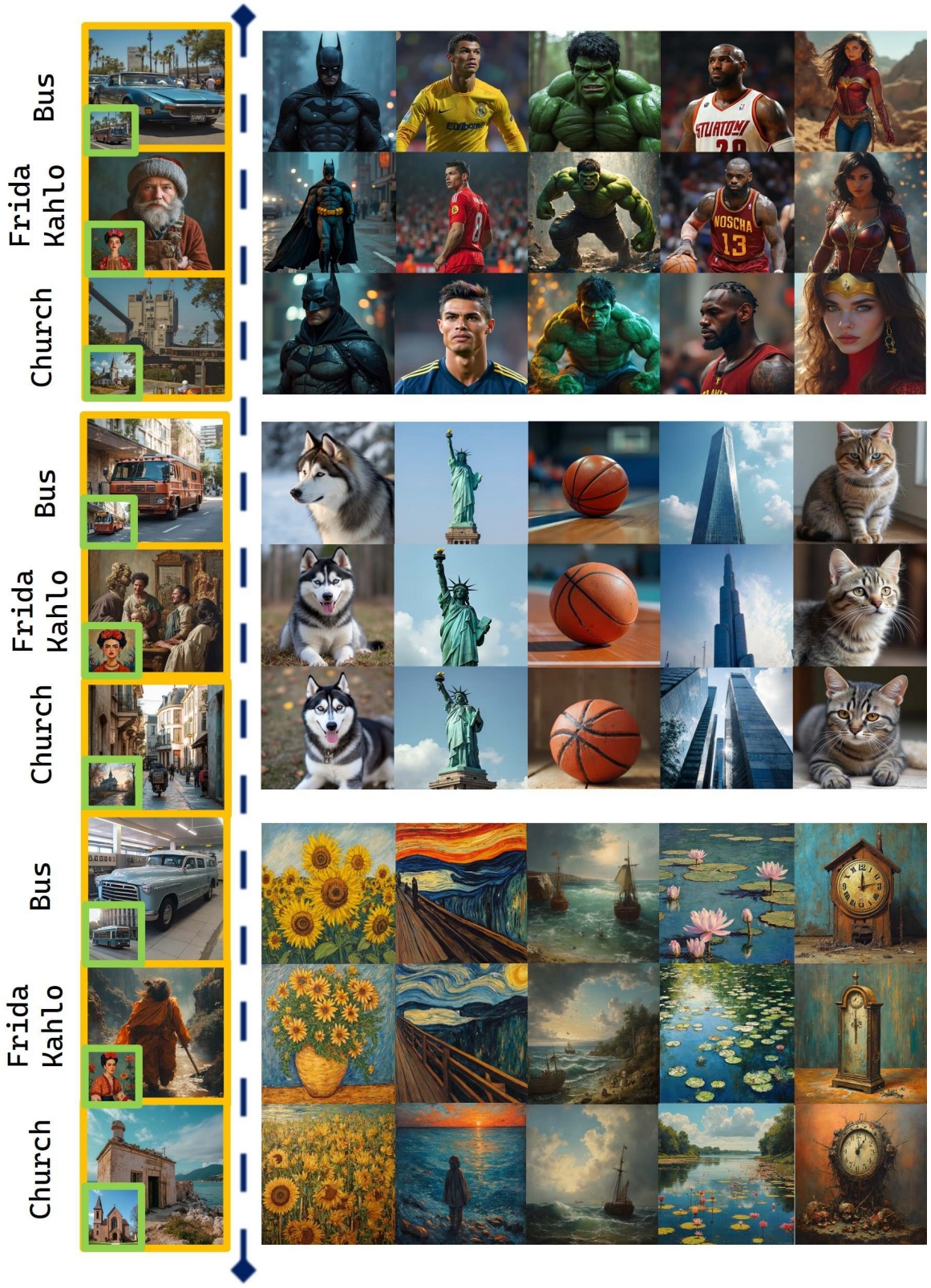

Figure 16. **Visualization on LoRA Disentanglement**. The left side of the blue dashed line delineates the erasure-concept-generated images (yellow box) and the original image (green box at the lower left). The right side illustrates the result on unrelated concepts upon incorporating the LoRA associated with the erased concept. Top rows: *Celebrity*; Mid rows: *Object*; Last rows: *Art*.

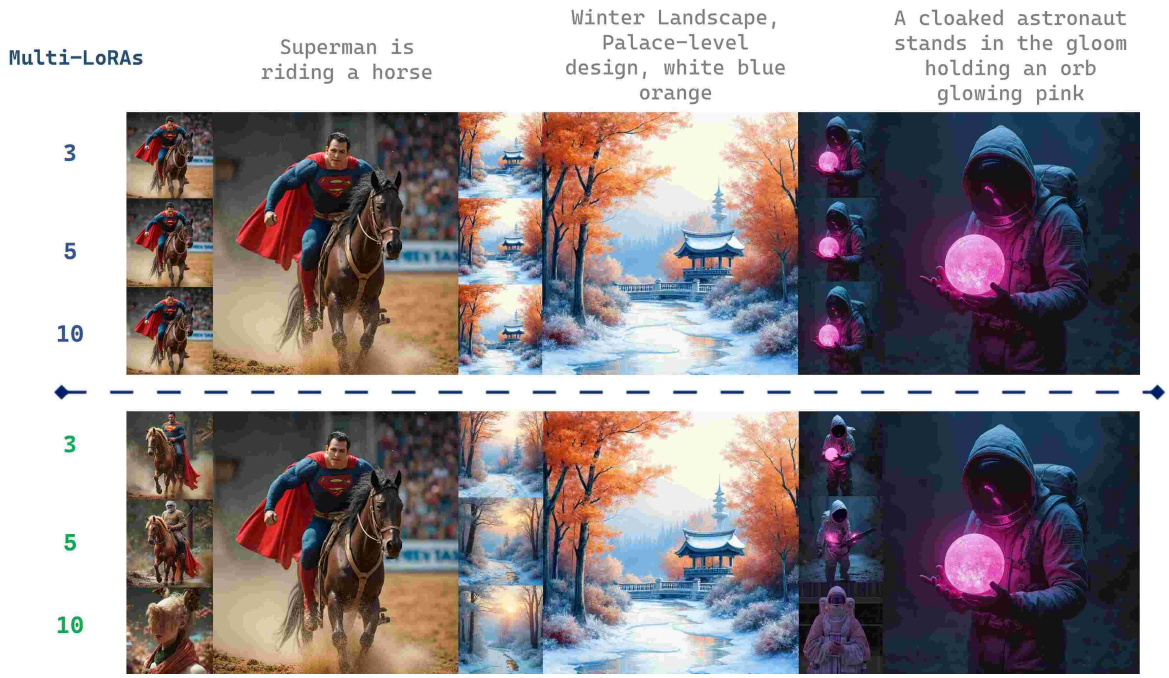

Figure 17. **Compositional LoRAs for irrelevant concepts**. We randomly sampled irrelevant concept-erased LoRAs and blending them in two ways: **Normalized Sum** (above the blue dotted line) and **Un-Normalized Sum** (below the blue dotted line).

Figure 18. **Compositional LoRAs for related concepts**. We find that through **Normalized Sum**, we can effectively erase multiple concepts at the same time.

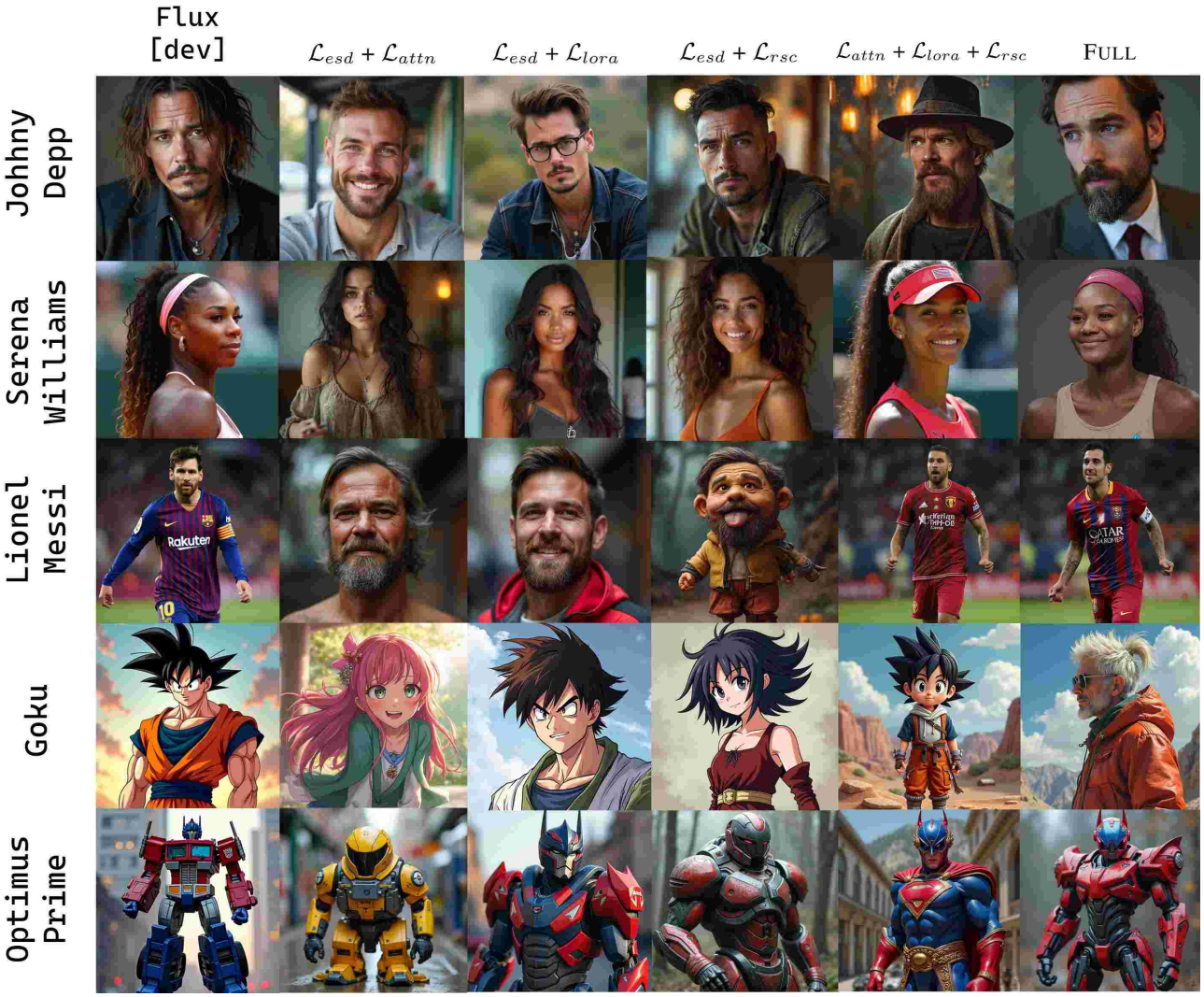

Figure 19. **Ablation Study on different loss configs**.