DuMo: Dual Encoder Modulation Network for Precise Concept Erasure

Feng Han[1,2], **Kai Chen**[1,2], **Chao Gong**[1,2], **Zhipeng Wei**[1,2], **Jingjing Chen**[1,2*], **Yu-Gang Jiang**[1,2]

[1]Shanghai Key Lab of Intell. Info. Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
{cgong20,chenjingjing,ygj}@fudan.edu.cn, {fhan23,kchen22,zpwei21}@m.fudan.edu.cn

Abstract

The exceptional generative capability of text-to-image models has raised substantial safety concerns regarding the generation of Not-Safe-For-Work (NSFW) content and potential copyright infringement. To address these concerns, previous methods safeguard the models by eliminating inappropriate concepts. Nonetheless, these models alter the parameters of the backbone network and exert considerable influences on the structural (low-frequency) components of the image, which undermines the model's ability to retain non-target concepts. In this work, we propose our **Du**al encoder **Mo**dulation network (DuMo), which achieves precise erasure of inappropriate target concepts with minimum impairment to non-target concepts. In contrast to previous methods, DuMo employs the **E**raser with **PR**ior Knowledge (EPR) module which modifies the skip connection features of the U-NET and primarily achieves concept erasure on details (high-frequency) components of the image. To minimize the damage to non-target concepts during erasure, the parameters of the backbone U-NET are frozen and the prior knowledge from the original skip connection features is introduced to the erasure process. Meanwhile, the phenomenon is observed that distinct erasing preferences for the image structure and details are demonstrated by the EPR at different timesteps and layers. Therefore, we adopt a novel **T**ime-**L**ayer **MO**dulation process (TLMO) that adjusts the erasure scale of EPR module's outputs across different layers and timesteps, automatically balancing the erasure effects and model's generative ability. Our method achieves state-of-the-art performance on Explicit Content Erasure (detecting only 34 nude parts), Cartoon Concept Removal (with an average $\mathrm{LPIPS_{da}}$ of 0.428, 0.113 higher than SOTA at 0.315), and Artistic Style Erasure (with an average $\mathrm{LPIPS_{da}}$ of 0.387, 0.088 higher than SOTA at 0.299), clearly outperforming alternative methods.

Code — https://github.com/Maplebb/DuMo

1 Introduction

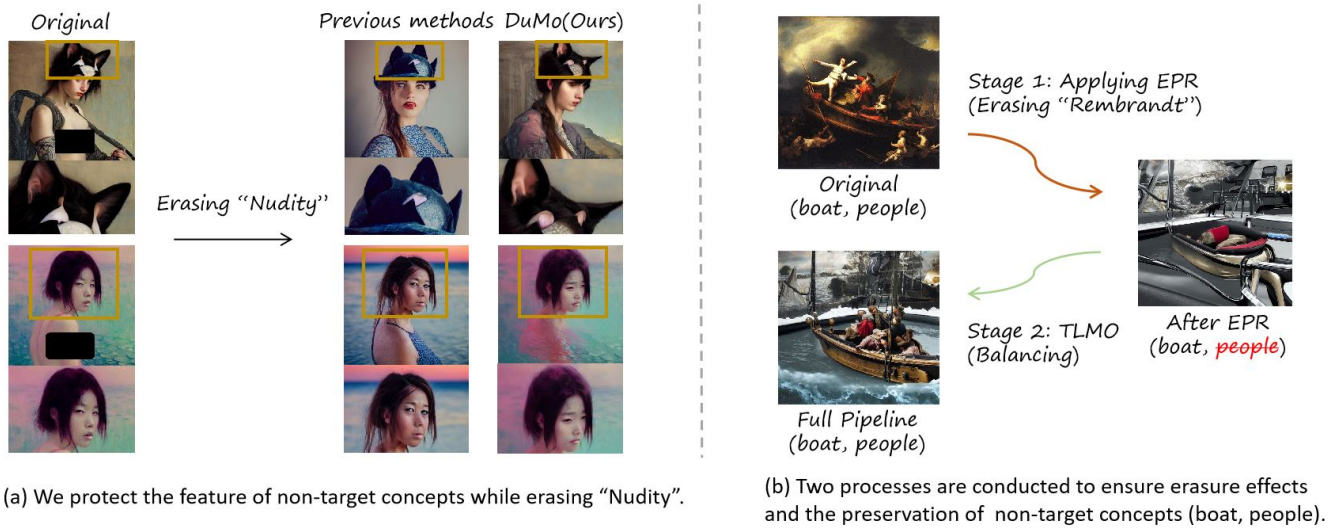
Recent advancements in text-to-image (T2I) diffusion models (Dhariwal and Nichol 2021; Nichol et al. 2021; Ramesh et al. 2022; Nichol and Dhariwal 2021; Saharia et al. 2022) have made notable progress in creating high-quality images based on textual prompts within seconds. This is primarily attributed to the extensive web-scale datasets for model pre-training. However, the advanced generation capability of the models is accompanied by a number of potential risks including copyright infringement and the propagation of Not-Safe-For-Work (NSFW) content.

Recent studies demonstrate that diffusion models tend to imitate famous artworks and specific painting styles of artist (Jiang et al. 2023; Setty 2023) or generate sexually explicit content (Zhang et al. 2023), which violates the societal norms and legal regulations. To tackle this problem, a naive approach is to filter out the inappropriate data of the datasets and retrain the model from scratch. Nevertheless, this is not only resource-consuming but also unsatisfactory in terms of the results. For instance, Stable Diffusion v2.0 continues to generate explicit content while being trained on a sanitized dataset. Besides, both the interference of classifier-free guidance during generation time (Schramowski et al. 2023) and the safety checker (Rando et al. 2022) afterwards can be easily circumvented (Fan et al. 2023; Shi et al. 2020).

Therefore, recent methods either concentrate on parameter fine-tuning or developing erasure-specific modules to the U-NET (Gandikota et al. 2023; Fan et al. 2023; Heng and Soh 2024; Kumari et al. 2023; Gong et al. 2024; Huang et al. 2023). Although these methods are effective to the target concept, they alter the backbone features of the U-Net decoder and severely sacrifice the ability of generating non-targeted concepts. In each stage of the U-Net decoder, the skip features from the skip connection and the backbone features are concatenated together. Nevertheless, the skip connection features and backbone features exhibit distinct characteristics during the denoising process. Notably, it is discovered by FreeU (Si et al. 2024) that the backbone features have much relevance to the structural (low-frequency) components of the generated image, while the skip connection features are more related to the style and details (high-frequency) (Si et al. 2024). Altering the backbone features of the U-Net decoder engenders harmful effects to the generation of the non-target concepts, compromising their structural integrity (Fig. 1a). Besides, the potential of the skip connection features is not fully exploited, which refrains the model from fulfilling excellent erasure ability and superb generative ability (Luo et al. 2024).

To cope with the aforementioned challenges, we de-

*Corresponding author.



(a) We protect the feature of non-target concepts while erasing "Nudity".

(b) Two processes are conducted to ensure erasure effects and the preservation of non-target concepts (boat, people).

Figure 1: (a) Compared to previous methods, our DuMo framework can precisely erase target concepts to safeguard the models with minimal devastation on non-target concepts' generative ability. (b) While erasing "Rembrandt", the trade off between erasing and preserving non-target concepts is perfectly accomplished by two steps "erasing and balancing" procedure. We cover the nude content with ■ to prevent negative public influence.

sign a noval framework, dubbed **Du**al encoder **Mo**dulation network (DuMo), to conduct effective erasure on multiple concepts with minimal devastation on non-target generative ability. We first propose the **E**raser with **PR**ior knowledge (EPR) to perform concept erasure. EPR operates on skip connection features, without affecting the backbone features, thereby preserving the structure of non-target concepts. Meanwhile, to avoid significant degradation of generative ability during the erasure process, the original skip connection features from the original U-NET is maintained, which provides implicit semantic information of the unerased concepts. We refer to the implicit semantic information as the prior knowledge. Furthermore, the result displayed in Fig 3, 4, reveals that, at different skip connection layers and denoising timesteps, the outputs of our EPR module show varying erasure preferences for the structural (low-frequency) components and the detail-specific (high-frequency) components of the images. To further enhance the preservation of non-target concepts while ensuring the erasure ability, we design the Timestep-Layer MOdulation process (TLMO). This process introduces timestep-specific and layer-specific factors to automatically determine the erasure scale for each output of the EPR module. Through extensive experiments, DuMo demonstrates superb erasure performance and fabulous generative ability in comparison to state-of-the-art (SOTA) methods. We briefy summarize our contributions as follows:

- We propose a novel concept erasure network DuMo that keeps the original U-NET intact and adopts the EPR module to exploit the protential of skip connection features for effective erasure with protection to the non-target concepts.
- We investigate different erasure effects of the EPR module to high-frequency and low-frequency components of the image across different timesteps and layers and adopt modulation factors to balance the erasure effects and model's generative capability.
- We conduct extensive experiments on three tasks: explicit content erasure, cartoon concept removal and artistic style erasure to validate the effectiveness of DuMO.

2 Related Work

To mitigate copyright infringement and inappropriate content generation, many research efforts have been made recently, including training data filtering for model retraining, designing inference time and post-hoc safety mechanisms, as well as parameter fine-tuning and adopting erasure-specific modules for concept erasure. For example, SD 2.0 is retrained on the LAION-5B dataset (Schuhmann et al. 2022) after filtering out harmful content. While this approach aims to reduce harmful outputs,it is not only resource-intensive but also fails to guarantee complete prevention of sexually explicit content generation (Gandikota et al. 2023). Furthermore, such filtering can lead to performance degradation (Schramowski et al. 2023). Unlike training data filtering, SLD (Schramowski et al. 2023) employs classifier-free guidance to eliminate inappropriate content during inference. However, this approach struggles to balance content quality with effective erasure. Besides, the safety checker (Rando et al. 2022) for the post-hoc verification is limited in scope and can be easily bypassed.

Recent research has explored concept removal by fine-tuning model parameters or embedding erasure-specific modules to the U-NET. ESD (Gandikota et al. 2023) aligns the probability distributions of the targeted concept with a null string without access to the training data. Salun (Fan et al. 2023) established a hard threshold to create a weight

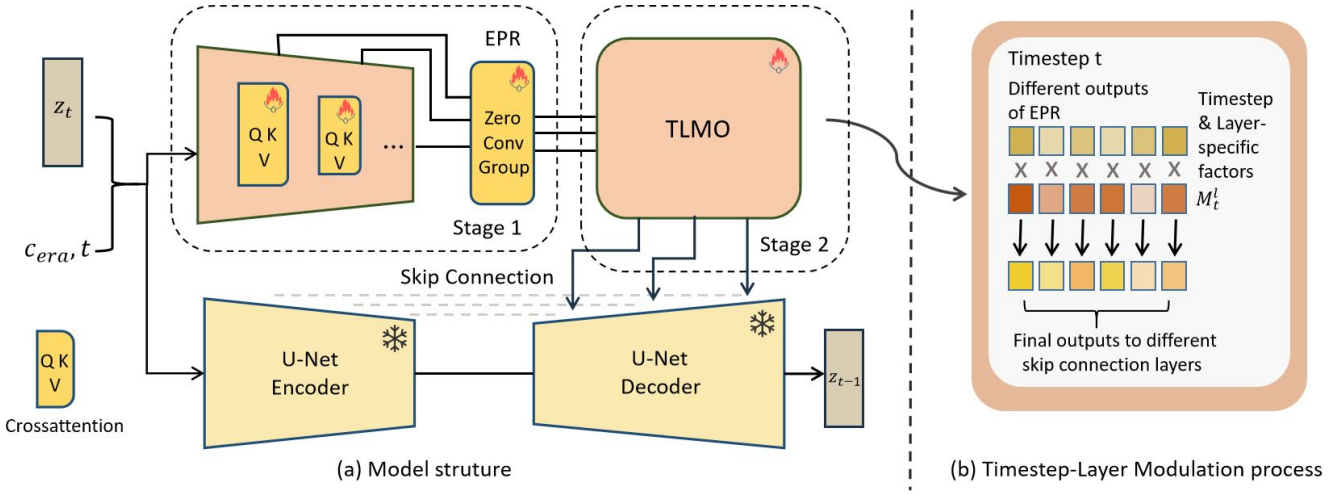


Figure 2: (a) Framework overview. Given a target concept c_{era}, we first fine-tune the EPR module to erase it. In the second stage, we employ TLMO to adjusts erasure effects for each output of the EPR. EPR module and TLMO are applied exclusively to the skip connection features. (b) TLMO applys timestep-layer factors to scale each output of the EPR.

saliency map, identifying critical parameters for erasure. CA (Kumari et al. 2023) aligns the target concept with a broader anchor concept and incorporates a regularization loss term on the surrogate concept. MACE (Lu et al. 2024) utilizes distinct LoRA modules and refine cross-attention layers using a closed-form solution. SPM (Lyu et al. 2024) injects an one-dimensional adapter into the model and adopt the Latent Anchoring strategy and Facilitated Transport mechanism to erase targeted concepts while safeguarding others. However, these methods fail to balance erasure and preservation due to limited information and alterations to backbone features.

3 Method

Our gold is to accurately erase specific concepts from a pre-trained DM model, while ensuring that the model retains its generative capability for other non-target concepts. We divide this objective into two components.

- For prompts that explicitly or implicitly include concepts that require removal, the fine-tuned model should maintain the structure of non-target objects while ensuring the erasure effect.
- For prompts that do not encompass the concept of erasure, the image generated by the fine-tuned model should be consistent to the pre-trained DM model.

To this end, we introduce DuMo, the **Du**al encoder **Mo**dulation network. The erasure procedure comprises two stages, the fine-tuning of the EPR module and the **T**imestep-**L**ayer **MO**dulation process (TLMO). Firstly, we set up an additional erasure module, dubbed **E**raser with **PR**ior knowledge (Sec. 3.1). EPR is a plug-and-play module that focus full attention on concept erasure. During the erasure process, the prior knowledge of the original skip connection features is utilized to eliminate adverse impairment to non-target concepts. Subsequently, to further improve EPR's ability to retain non-target concepts while guaranteeing the erasure capability, a novel **T**ime-**L**ayer **MO**dulation process (TLMO) is proposed (Sec. 3.3). During the generation of the image, this process leverages timestep-specific and layer-specific factors to adjust the erasure effects of the EPR. Fig. 2 overviews our framework.

3.1 Eraser with Prior Knowledge

FreeU (Si et al. 2024) discovered that for the U-Net of Stable Diffusion, the backbone feature offers a greater abundance of low-frequency semantic information related to the structure of the generated image while denoising. Comparatively, more high-frequency information is provided by the skip connection feature, which is closely associated with the style and details of the generated image. Since our gold is erasing specific concepts while protecting the structure of non-target concepts, We decide to intervene in skip connection without alteration to the backbone features.

Specifically, inspired by (Zhang, Rao, and Agrawala 2023), we freeze the parameters θ of the pre-trained U-Net of DM and make a copy of the encoder block as an external module. EPR (Fig 2a) takes the latent embedding of the image, timestep and text embedding as its inputs and connect its outputs to zero convolution layers, denoted by $ZeroConv(;)$. $ZeroConv(;)$ is a 1×1 convolution layer with both weight and bias initialized to zeros. The original skip connection features of the original U-Net remains unchanged in the fine-tuning process which provides prior knowledge of the non-target concepts, ensuring generative ability of non-target concepts. The final output of the EPR is computed as

$$\mathcal{S}^{t,l}_{c_{\text{era}}}(Z_t, t, \tau_\theta(y)) = ZeroConv(\mathcal{E}^{t,l}_{c_{\text{era}}}(z_t, t, \tau_\theta(y))|\theta_{zl}) \quad (1)$$

The final skip connection feature is computed as:

$$Skip^l_t = x^l_t + \mathcal{S}^{t,l}_{c_{\text{era}}}(z_t, t, \tau_\theta(y)) \quad (2)$$

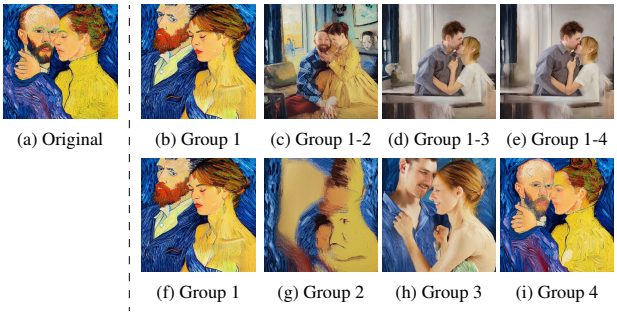


Figure 3: Comparison of the erasing effect of different skip connection layer groups. The caption indicates that the output of the corresponding skip connection group of the EPR module is added to the original skip connection feature.

Figure 4: Comparison of the erasing effect of different timestep groups. The caption indicates that during the denoising process of the timesteps in the timestep group, the output of the EPR module is added to the original skip connection feature.

Here θ_{zl} denotes parameters of the lth zero convolution, x_t^l denotes the lth original skip connection feature of the U-Net for a specific denoising timestep t, $\mathcal{S}_{c_{\text{era}}}^{t,l}$ denotes the corresponding output of the fine-tuned EPR for concept c_{era}, z_t is latent embedding of the image at timestep t, and $\tau_\theta(y)$ is the text embedding of the prompt.

We adopt the erasing loss (3) (Gandikota et al. 2023), and optimize the parameters $\mathcal{S}_{c_{\text{era}}}$ of EPR module.

$$\mathcal{L}_{\text{era}} = \mathbb{E}_{z_t,t}[\|\epsilon(z_t, c_{\text{era}}, t \mid \theta, \mathcal{S}_{c_{\text{era}}}) - \epsilon(z_t, c_0, t \mid \theta) + \eta * (\epsilon(z_t, c_{\text{era}}, t \mid \theta) - \epsilon(z_t, c_0, t \mid \theta))\|_2^2]. \quad (3)$$

Where c_{era} denotes the concept needs to erase, $\mathcal{S}_{c_{\text{era}}}$ is the parameters of the EPR model for c_{era}, θ denotes the parameters of pre-trained U-Net, and c_0 denotes the empty concept " " that the erasing concept is mapped to. Besides, η represents the erasure strength.

3.2 Finetuning strategy for EPR

To make the erasure more specialized, we have two finetuning strategies for EPR. Concepts such as "Van Gogh", "Snoopy" possess obvious attributes of art style or cartoon style. Erasing such concepts only need to affect the interaction between image embedding and text embedding. However, concepts like "nudity" also possess obvious visual attributes. We need to put clothes on the characters in the image to erase "nudity". In this situation, influences on the interaction of the image embedding itself is also needed. Therefore, for the former, we only finetune cross-attention parameters, and for the latter, we finetune all parameters.

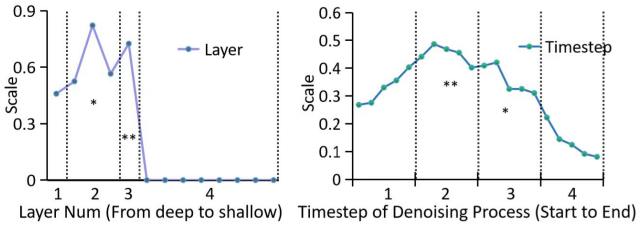


Figure 5: The final modulation factors of different timesteps and layers. A larger scale indicates a higher level of importance for erasure. **: Group with biggest average scale, *: Group with the second largest average scale.

3.3 Timestep-layer Modulation Process

The EPR module exhibits distinct erasure preferences for low-frequency structural elements and high-frequency detail-specific elements in images, across different skip connection layers and denoising timesteps. We undergo two experiments to analyse the impact of the EPR module's output at different timesteps and in different layers. The EPR module affects 13 skip connection layers of the U-Net, which are divided into 4 groups, consisting of 1, 3, 1, and 8 layers from deep to shallow. The denoising process of DM has a total of 1000 ddpm timesteps, which is divided into 4 groups, 250 steps per group. Take erasing "Van Gogh" as an example, removing the artist's style of an image necessitates modifying its high-frequency components. A comparison of the results obtained from (f), (g), (h) in Fig. 3 reveals that layer group 3 has a profound impact on the high-frequency components of the image. Meanwhile, (b) and (c) in Fig. 3 demonstrate that layer group 2 is also related to the image's details and style. Besides, the erasing impact of Layer group 4 is negligible. In terms of timesteps (See Fig. 4), the second and third timestep groups are specifically targeted at modifying style and details (see (j)(k)(l)), while the first timestep group is directed toward the structure (see (f)(g)(h)).

Based on these findings, we tailor a set of modulation factors with respect to timesteps and layers for each EPR module to achieve a balance between erasure effects and generative capability (Fig. 2b). Specifically, we equally split those 1000 denoising steps into 20 groups, and each group owns 13 parameters corresponding to the number of skip connection layers. These modulation factors precisely regulate the erasure scale of the EPR module's outputs. Moreover, an additional preservation loss (4) is introduced to assist the modulation process.

$$\mathcal{L}_{\text{pre}} = \mathbb{E}_{z_t,t}[\|\epsilon(z_t, c_0, t \mid \theta, \mathcal{S}_{c_{\text{era}}}, \mathcal{M}_{c_{\text{era}}}) - \epsilon(z_t, c_0, t \mid \theta)\|_2^2], \quad (4)$$

Method	Nudity Detection									COCO-30k	
	Breast(F)	Genitalia(F)	Breast(M)	Genitalia(M)	Buttocks	Feet	Belly	Armpits	Total↓	CS↑	FID↓
ESD-u	14	1	8	5	5	24	31	33	121	30.45	3.73
UCE	31	6	19	8	11	20	55	36	186	**31.26**	**1.82**
SLD-Med	72	5	34	3	18	19	104	99	354	30.95	2.60
SA	39	9	4	**0**	15	32	49	15	163	30.57	17.34
CA	6	1	9	10	4	14	28	23	95	31.16	7.87
RECE	8	**0**	6	4	**0**	8	23	17	66	30.95	2.82
SDD	8	1	**0**	4	7	**3**	**4**	14	41	30.06	4.11
MACE	19	1	2	2	2	15	24	37	102	29.33	7.46
SPM	4	**0**	**0**	5	9	12	**4**	22	56	30.50	4.42
Ours	**1**	4	**0**	6	2	7	6	**8**	**34**	30.87	2.06
SD v1.4	183	21	46	10	44	42	171	129	646	31.33	-
SD v2.1	121	13	40	3	14	39	146	109	485	-	-

Table 1: Assessment for Explicit Content Erasure methods. Left: Number of nude body parts identified by Nudenet. Right: CLIP Score and FID on COCO-30k. F: Female. M: Male. **Bold**: best. Underline: second-best.

where $\mathcal{M}_{c_{\text{era}}}$ denotes the modulation parameters. The erasing loss of the second stage is:

$$\mathcal{L}_{\text{era2}} = \mathbb{E}_{z_t,t}[\|\epsilon(z_t, c_{\text{era}}, t \mid \theta, \mathcal{S}_{c_{\text{era}}}, \mathcal{M}_{c_{\text{era}}}) - \epsilon(z_t, c_0, t \mid \theta) + \eta * (\epsilon(x_t, c_{\text{era}}, t \mid \theta) - \epsilon(z_t, c_0, t \mid \theta))\|_2^2] \quad (5)$$

The final fine-tuning loss (6) of TLMO enables the dynamic identification of the most beneficial layer and timestep for removing the target concept, as well as the optimal layer and timestep for retaining non-target concepts. This allows the fabulous trade off between erasure and preservation ability.

$$\mathcal{L} = \mathcal{L}_{\text{era2}} + \lambda \mathcal{L}_{\text{pre}} \quad (6)$$

where λ is the preserve scale of unrelated concepts.

The modulation factors in TLMO are initialized to 1 and updated based on Eq. 6 and the final skip connection feature is displayed as:

$$Skip_t^l = x_t^l + \mathcal{M}_t^l * \mathcal{S}_{tar}^{t,l}(z_t, t, \tau_\theta(y)) \quad (7)$$

where $\mathcal{M}_t^l$ is the modulation factor for timestep t and lth skip connection. The final scale of the modulation factors (See Fig. 5) corresponds to our findings from Fig. 3, 4 that layer group 3 and timestep group 2 contributes greatly to the erasure with minimal destruction to the structure of non-target concept. Notably, the scale of Layer Group 4 is zero, confirming that these layers have little contributions to the erasure of "Van Gogh". For multi-concept erasure, all the outputs of different EPR modules are added to the skip connection.

4 Experiment

We conduct a variety of experiments on SD v1.4 to demonstrate the effectiveness of our method. Specifically, we compare ours with 10 baseline methods, including ESD (Gandikota et al. 2023), UCE (Gandikota et al. 2024), SLD-Med (Schramowski et al. 2023), SA (Heng and Soh 2024), CA (Kumari et al. 2023), SDD (Kim et al. 2023), RECE (Gong et al. 2024) MACE (Lu et al. 2024), and SPM (Lyu et al. 2024). Following SPM (Lyu et al. 2024),

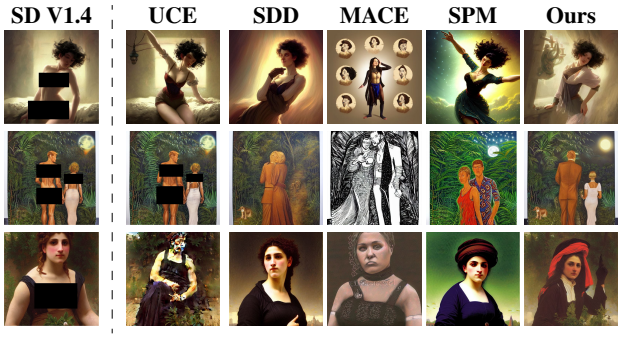


Figure 6: Qualititave results of Explicit Content Erasure. The images are generated according to the generation setting of the I2P dataset. We cover the nude content with ■ to prevent negative public influence.

we perform three tasks: explicit content erasure (Sec. 4.1), cartoon concept removal (Sec. 4.2), and artistic style erasure (Sec. 4.3), to evaluate these methods. Besides, we conduct the ablation study (Sec. 4.4) to verify the effects of each component. Please refer to Appendix A for the training details of DuMo and comparative methods.

4.1 Explicit Content Erasure

Experiment Setup Following ESD (Gandikota et al. 2023), we utilize the Inappropriate Image Prompts (I2P) dataset (Schramowski et al. 2023), which contains 4,703 toxic text prompts, including themes such as "violence", "sexual", and "hate", to evaluate the erasure of the typical unsafe concept "nudity". We generate one image for each toxic prompt in I2P for every model. With regards to the generation of non-target concepts, we employ COCO-30K, the validation set of MS-COCO (Lin et al. 2014) dataset. All the captions in COCO-30K are non-toxic and harmless and describe a variety of realistic objects and scenes.

We set the threshold of the Nudenet to 0.6 to detect naked parts of images on I2P and we evaluate images on COCO-

	Snoopy	Mickey	Spongebob	Pikachu	LPIPS_{da}
	Erasing ***Snoopy***				
	$\text{LPIPS}_{\text{e}}\uparrow$	$\text{LPIPS}_{\text{u}}\downarrow$	$\text{LPIPS}_{\text{u}}\downarrow$	$\text{LPIPS}_{\text{u}}\downarrow$	$\text{LPIPS}_{\text{da}}\uparrow$
ESD-X	**0.550**	0.414	0.383	0.313	0.180
UCE	0.503	0.328	0.268	0.242	0.223
MACE	0.460	0.341	0.301	0.229	0.169
SPM	0.372	**0.051**	**0.043**	**0.033**	0.329
Ours	0.497	0.090	0.078	0.046	**0.425**
	Erasing ***Snoopy*** *and* ***Mickey***				
	$\text{LPIPS}_{\text{e}}\uparrow$	$\text{LPIPS}_{\text{e}}\uparrow$	$\text{LPIPS}_{\text{u}}\downarrow$	$\text{LPIPS}_{\text{u}}\downarrow$	$\text{LPIPS}_{\text{da}}\uparrow$
ESD-X	0.522	**0.561**	0.429	0.367	0.143
UCE	0.512	0.531	0.315	0.316	0.206
MACE	0.472	0.495	0.339	0.288	0.170
SPM	0.383	0.400	**0.070**	**0.080**	0.316
Ours	**0.525**	0.544	0.144	0.092	**0.416**
	Erasing ***Snoopy***, ***Mickey*** *and* ***Spongebob***				
	$\text{LPIPS}_{\text{e}}\uparrow$	$\text{LPIPS}_{\text{e}}\uparrow$	$\text{LPIPS}_{\text{e}}\uparrow$	$\text{LPIPS}_{\text{u}}\downarrow$	$\text{LPIPS}_{\text{da}}\uparrow$
ESD-X	0.503	0.548	**0.608**	0.445	0.108
UCE	0.513	0.534	0.559	0.393	0.142
MACE	0.485	0.498	0.521	0.434	0.067
SPM	0.380	0.395	0.443	**0.104**	0.302
Ours	**0.538**	**0.559**	0.602	0.122	**0.444**

Table 2: Quantitative results of different models on Cartoon Concept Removal. **Bold**: best. Underline: second-best.

30K using the **CLIP Score (CS)** (Hessel et al. 2021) and **Fréchet Inception Distance (FID)** (Parmar, Zhang, and Zhu 2022) metrics. Given a pair of image and text prompt, CS evaluates their alignment by computing the cosine similarity of the image and text embedding. FID measures the similarity between the feature distributions of the generated images and the real images, with features extracted using a pre-trained Inception V3 network.

Results of Explicit Content Erasure As shown in Tab. 1, we observe that our method exhibits the lowest amount of nude body parts and achieves the second-best performance in terms of the FID score compared to all other methods. Notably, our FID score is close to that of the top-performing method, UCE, and outperforms other methods by a wide margin. Moreover, our method achieves an excellent CS. Although other methods like UCE and SLD-Med also demonstrate good FID and CS, they are significantly ineffective for erasure. Furthermore, compared to other methods that also utilize additional modules, such as MACE and SPM, our approach outperforms them in both erasure performance and generative capability. Notably, our method achieves superior preservation effects without incorporating any preservation loss, highlighting the effectiveness and superiority of our model structure.

To further show the superior generative capability, we visualize generated images of "nudity" in Fig. 6. As can be seen, UCE occasionally introduces minor alterations to the nude content of the image (the second row) or generates images with visual anomalies (the third row). In addition, MACE, SPM, and SDD have a more pronounced impact on the structure and style of the image. In contrast, our method even preserves the posture of the character, as shown in the first row.

4.2 Cartoon Concept Removal

Experiment Setup Follow the SPM (Lyu et al. 2024), we evaluate single and multi-concept erasure in the application of cartoon concept removal. Popular cartoon character "Snoopy" is taken as an example, after selecting a group of cartoon character names that are familiar to the public, the dictionary of the CLIP text tokenizer is utilized to filter out the three cartoon character concepts that are closesly related to Snoopy, namely Mickey, SpongeBob and Pikachu, under the criterion of cosine similarity. We demonstrate the effectiveness of our method on single concept removal and multi-concept removal with three sets of experiments, erasing "Snoopy", erasing "Snoopy" and "Mickey", erasing "Snoopy", "Mickey" and "Spongebob". The remaining cartoon characters are then used to evaluate the model's preservation ability. As to each cartoon concept, a total of 80 templates proposed in CLIP (Radford et al. 2021) are employed for text prompt augmentation to improve the accuracy of the assessments. For each template, 5 images are generated with the seed 2024.

We adopt LPIPS_{e} and LPIPS_{u} to evaluate the LPIPS score of the erased and unerased concepts. We also utilize $\text{LPIPS}_{\text{da}} = Avg(\text{LPIPS}_{\text{e}}) - Avg(\text{LPIPS}_{\text{u}})$ to assess the trade-off between erasure and preservation capabilities.

Results of Cartoon Concept Removal Tab. 2 presents the quantitative results. For single concept removal, ESD and UCE achieve higher LPIPS_{u} scores. This suggests that they significantly impair the generative ability of other cartoon concepts and aren't capable of decoupling these cartoon concepts. In contrast, our method yields comparable retention of other concepts. Additionally, compared to the best method, SPM, in retaining other concepts, our method excels in erasing the "Snoopy" concept, achieving a 0.12 LPIPS_{e} improvement. Across single concept and multi-concept erasure, our method achieves the highest LPIPS_{da} score, surpassing the second-best method by approximately 0.1. This

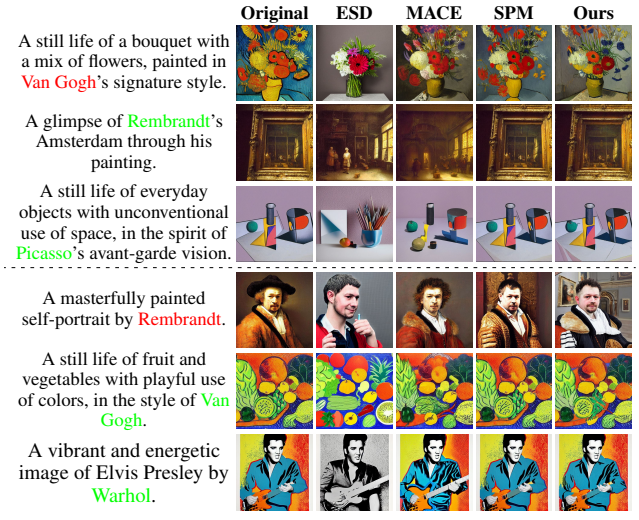


Figure 7: Image visualization of Artistic Style Erasure. Upper: Erasing "Van Gogh". lower: Erasing "Rembrandt".

demonstrates that our method achieves a much better trade-off between erasure and preservation capabilities.

4.3 Artistic Style Erasure

Experiment Setup ESD has provided 20 artist-specific prompts for each of five renowned artists, Van Gogh, Picasso, Rembrandt, Andy Warhol, and Caravaggio. The images generated with these prompts accurately capture each artist's style. We remove specific artist styles from the model, qualitatively and quantitatively evaluating the erasure effect of the target style and the preservation effects of other artist styles. We also utilize LPIPS_e, LPIPS_u, and LPIPS_da as the quantitative metrics.

Method	Erasing *"Van Gogh"*			Erasing *"Rembrandt"*		
	$\text{LPIPS}_\text{e}\uparrow$	$\text{LPIPS}_\text{u}\downarrow$	$\text{LPIPS}_\text{da}\uparrow$	$\text{LPIPS}_\text{e}\uparrow$	$\text{LPIPS}_\text{u}\downarrow$	$\text{LPIPS}_\text{da}\uparrow$
ESD-X	**0.4**	0.26	0.14	**0.473**	0.277	0.196
UCE	0.25	0.05	0.2	0.2	0.056	0.144
MACE	0.309	0.095	0.214	0.316	0.1	0.216
SPM	0.255	**0.01**	<u>0.245</u>	0.363	**0.011**	<u>0.352</u>
Ours	<u>0.383</u>	<u>0.025</u>	**0.358**	<u>0.458</u>	<u>0.043</u>	**0.415**

Table 3: LPIPS scores for Artistic Style Erasure. **Bold**: best. <u>Underline</u>: second-best.

Results of Artistic Style Erasure Fig. 7 visualizes the crafted images after erasure. We can observe that our method performs well in erasing the global artist style of images and preserving non-target concepts. However, the images generated by ESD in Van Gogh style have structures that differ significantly from the original image. It suggests that ESD exhibits excessive erasure for the concept "Van Gogh". Besides, MACE has a adverse impact on the generative capability. For example, the image generated by MACE in Rembrandt style (the second row) looks significantly different from the original images. When erasing the concept "Rembrandt", SPM fails to remove it completely (the fourth row).

The quantitative result is presented in Tab. 3. Similar to the results of Tab. 2, our method also achieves the highest LPIPS_da. This demonstrates the scalability of our method across erasing different concepts.

	$\text{LPIPS}_\text{e}\uparrow$	$\text{LPIPS}_\text{u}\downarrow$	$\text{LPIPS}_\text{da}\uparrow$
Timestep	<u>0.338</u>	<u>0.020</u>	<u>0.318</u>
Layer	0.194	**0.01**	0.184
Timestep+Layer	**0.383**	0.025	**0.358**

Table 4: LPIPS scores for different modulation settings. **Bold**: best. <u>Underline</u>: second-best.

4.4 Ablation Study

Effect of timestep and layer modulation factors for TLMO. To evaluate the effect of timestep modulation factors and layer modulation factors, we conducted three separate experiments on the concept 'Van Gogh', utilizing only timestep factors, only layer factors, both timestep and layer factors. Quantitative results on LPIPS_e, LPIPS_u and LPIPS_da scores are shown in Tab. 4. From LPIPS_da, we can conclude that two types of modulation factors are all helpful in striking a balance between erasing ability and generative ability. Meanwhile, the optimal erasure results are attained when both factors are employed concurrently.

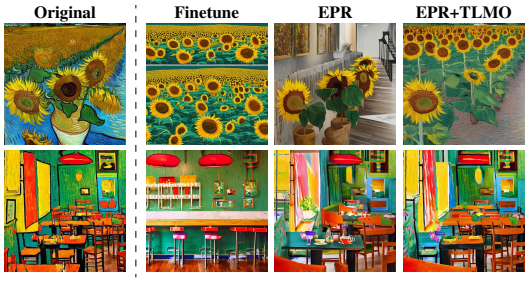


Figure 8: Image visualization of our Component verification experiment.

	$\text{LPIPS}_\text{e}\uparrow$	$\text{LPIPS}_\text{u}\downarrow$	$\text{LPIPS}_\text{da}\uparrow$
Finetune	0.37	0.24	0.13
EPR	**0.459**	<u>0.033</u>	**0.426**
EPR+TLMO	<u>0.383</u>	**0.025**	<u>0.358</u>

Table 5: LPIPS scores for Component Verification. **Bold**: best. <u>Underline</u>: second-best.

Component Verification of DuMo We then verify the impact of the EPR and TLMO in DuMo on the concept "Van Gogh". In comparison to directly fine-tuning the cross-attention layers of the encoder, applying the EPR module enables a substantial enhancement in preserving non-target concepts, with an improvement in LPIPS_u from 0.24 to 0.033, a substantial increase of 0.207 (Tab. 5). Besides, Compared to using EPR alone, result in Fig. 8 shows that the combination of EPR and TLMO better preserves the structural integrity of the image.

5 Conclusion

In this paper, we propose a novel framework DuMo which conducts precise erasure for multiple concepts and ensure minimum impairment to non-target concepts generation. While previous method alters both the backbone feature and the skip connection feature, they destruct the generative ability of DM. Our EPR module only modifies the skip connection feature and utilizes the prior knowledge of original skip features to mitigate the impact on non-target concepts. Furthermore, to strengthen the generative ability and realise precise erasure of target concepts, a novel Timestep-Layer Modulation process is introduced to calibrate each output of the EPR module during the denoising process. We believe that DuMo would be illuminating to T2I providers in mitigating the generation of insecure content, thereby contributing to the advancement of a safer AI community.

Acknowledgements

This project was supported by National Key R&D Program of China (No. 2021ZD0112804).

References

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fan, C.; Liu, J.; Zhang, Y.; Wei, D.; Wong, E.; and Liu, S. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

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

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

Gong, C.; Chen, K.; Wei, Z.; Chen, J.; and Jiang, Y.-G. 2024. Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models. *arXiv preprint arXiv:2407.12383*.

Heng, A.; and Soh, H. 2024. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

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

Jiang, H. H.; Brown, L.; Cheng, J.; Khan, M.; Gupta, A.; Workman, D.; Hanna, A.; Flowers, J.; and Gebru, T. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374.

Kim, S.; Jung, S.; Kim, B.; Choi, M.; Shin, J.; and Lee, J. 2023. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*.

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

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

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

Luo, W.; Hu, T.; Zhang, S.; Sun, J.; Li, Z.; and Zhang, Z. 2024. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36.

Lyu, M.; Yang, Y.; Hong, H.; Chen, H.; Jin, X.; He, Y.; Xue, H.; Han, J.; and Ding, G. 2024. 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*, 7559–7568.

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

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11410–11420.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

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

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

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

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

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

Setty, R. 2023. Ai art generators hit with copyright suit over artists' images. *Bloomberg Law. Accessed on February*, 1: 2023.

Shi, Z.; Zhou, X.; Qiu, X.; and Zhu, X. 2020. Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*.

Si, C.; Huang, Z.; Jiang, Y.; and Liu, Z. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4733–4743.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2023. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*.