Multimodal Robustness Benchmark for Concept Erasure in Diffusion Models

Ju Hsuan Weng[1,2] **Jia-Wei Liao**[1,2] **Cheng-Fu Chou**[1] **Jun-Cheng Chen**[2]
[1] National Taiwan University,
[2] Research Center for Information Technology Innovation, Academia Sinica
{r12922a05, d11922016}@csie.ntu.edu.tw

Abstract

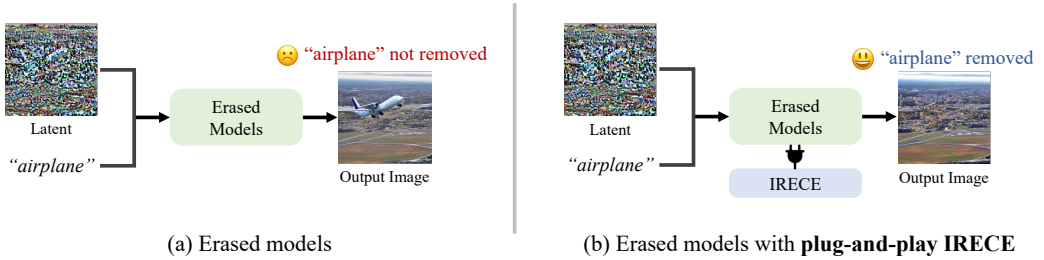
Text-to-image diffusion models may produce harmful or copyrighted content, motivating research on concept erasure. However, existing approaches mainly target text prompts, overlooking other input modalities crucial to real-world applications such as image editing and personalization. These modalities can act as attack surfaces where erased concepts reappear. To address this, we introduce a multimodal evaluation framework that benchmarks concept erasure methods across text prompts, learned embeddings, and inverted latents. Our analysis shows that current methods perform well on text prompts but largely fail under learned embeddings and latent inversion, with Concept Reproduction Rate (CRR) exceeding 90% in white-box settings. We further propose Inference-time Robustness Enhancement for Concept Erasure (IRECE), a plug-and-play module that localizes target concepts via cross-attention and perturbs their latents during denoising. Experiments show that IRECE restores robustness, reducing CRR by up to 40% under the most challenging white-box latent inversion while preserving visual quality.

1 Introduction

The contemporary diffusion models [1, 2, 3, 4, 5] have demonstrated remarkable progress in high-quality and versatile content generation, supporting tasks such as image synthesis, image editing [6, 7, 8], personalized generation [9, 10], and style transfer [11, 12]. However, training on large-scale uncurated datasets makes them prone to reproducing copyrighted [13, 14] or inappropriate content [15]. Retraining on filtered datasets offers a direct solution, but it is costly and often degrades generative quality [16]. Recent studies therefore explores **concept erasure** [17, 18, 19, 20, 21, 22, 23, 24], which aims to prevent text-to-image diffusion models from generating harmful or copyrighted content by suppressing specific concepts through fine-tuning cross-attention layers of diffusion models without retraining from scratch.

Despite their effectiveness under text prompts, existing methods reveal critical weaknesses. In practice, users often rely on learned embeddings from personalization techniques [9, 10] or noisy latent from inversion methods [25, 26], which fall outside the assumptions of text-based concept erasure. Our analysis shows that while suppression is reliable for basic prompts, concepts frequently re-emerge with learned embeddings or inverted latents, reaching Concept Reproduction Rate (CRR) over 90% in white-box setting with unconditional prompt. Geroge *et al.* [27] also found that the erased concepts can be revived through fine-tuning the model with a few samples. This indicates that current methods primarily disrupt text–image alignment rather than fully removing concepts. Moreover, although adversarial prompt attacks [28, 29] expose vulnerabilities, defenses based on adversarial training [22] remain limited to textual inputs, underscoring the need to explore robustness beyond the text space. This raises our central research question: *How robust are concept erasure methods across different input modalities, and can their vulnerabilities be mitigated without retraining?*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.



(a) Erased models

(b) Erased models with **plug-and-play IRECE**

Figure 1: **Illustration of concept erasure under latent inversion**. Existing methods cannot effectively remove the target concept (*airplane*), whereas our IRECE method successfully suppresses it while preserving visual quality.

Motivated by these limitations, we introduce a novel multimodal evaluation benchmark that systematically evaluates the robustness of the concept erasure methods across three representative input settings for text-to-image diffusion models: text prompts, learned embeddings, and latent inversion, under both white-box and gray-box settings, respectively. With the extensive evaluations from this framework, the results demonstrate that the state-of-the-art concept erasure methods are still vulnerable to various inference-time attack methods. Building upon these insights, we further propose Inference-time Robustness Enhancement for Concept Erasure (IRECE), a plug-and-play module that localizes target concepts via cross-attention and selectively perturbs associated latents during denoising without retraining.

2 Multimodal Evaluation Framework

Most concept-erasure methods are evaluated only on text prompts. However, there exists multiple ways that the users can interact with the model beyond text prompts, such as learned embeddings, inverted latents, etc. Therefore, we design a comprehensive multimodal evaluation framework with three distinct settings:

Text Prompt Evaluation. First, we evaluate the erased models using both basic prompts containing the target concept and adversarial prompts generated following Ring-A-Bell [30].

Learned Embedding Evaluation. We adopt Textual Inversion (TI) [9] to encode visual concepts from the reference images into learned embeddings, which are then used to guide the erased model. Then, we evaluate three settings: **(i) White-box**, where TI is trained directly on the erased model; **(ii) Gray-box**, where TI is trained on a standard diffusion model without modifications; and **(iii) Gray-box with perturbations**, where small perturbations are added to reference images in the gray-box case to induce semantic shifts in the learned embeddings.

Inverted Latent Evaluation. We leverage DDIM inversion [31] to map a reference image into the initial noisy latent, later combined with the text prompt during sampling. This setup evaluates whether erased concepts re-emerge when the generative process is directly initialized from a concept-containing latent. We consider both **(i) White-Box inversion**, where inversion is performed directly on the erased model and **(ii) Gray-Box inversion** where inversion is performed on the standard diffusion model without modifications. Moreover, we adopt four prompt-pair strategies (Table 1 in the Appendix), covering the unconditional, generic, coarse, and explicit target cases.

3 Inference-time Robustness Enhancement for Concept Erasure (IRECE)

We propose *Inference-time Robustness Enhancement for Concept Erasure (IRECE)*, a plug-and-play module that improves the reliability of erased models without retraining. The core idea is to disrupt latent regions encoding the erased concept during inference while preserving the rest of the image. Starting from the initial noisy latent $\boldsymbol{x}_T$, the erased model θ_{era} progressively denoises it under the guidance of the sample prompt embedding $\boldsymbol{c}_{\text{sam}}$. At an intervention step t^*, we identify spatial regions linked to the target concept $\boldsymbol{c}_{\text{tgt}}$ using cross-attention maps A^{ℓ}_{cross} from each layer ℓ of the standard

model θ_{std}. These maps are upsampled to a common resolution and aggregated:

$$\boldsymbol{A} = \sum_{\ell=1}^{L} \text{Upsample}\left(A^{\ell}_{\text{cross}}(\boldsymbol{x}_t, \boldsymbol{c}_{\text{tgt}}; \theta_{\text{std}})\right),$$

A binary mask $\boldsymbol{M}$ is obtained by thresholding $\boldsymbol{A}$ with parameter τ, marking pixels most associated with the erased concept:

$$\boldsymbol{M}(i,j) = \begin{cases} 1, & \boldsymbol{A}(i,j) \geq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

The masked latent regions are then replaced with Gaussian noise $\boldsymbol{\xi}_t$:

$$\boldsymbol{x}_t^* = (\mathbf{1} - \boldsymbol{M}) \odot \boldsymbol{x}_t + \boldsymbol{M} \odot \boldsymbol{\xi}_t,$$

Denoising resumes from $\boldsymbol{x}_t^*$ according to

$$\boldsymbol{x}_{t-1} = \begin{cases} \text{DDIMStep}(\boldsymbol{x}_t^*, \boldsymbol{c}_{\text{sam}}, t, \theta_{\text{era}}), & \text{if } t = t^*, \\ \text{DDIMStep}(\boldsymbol{x}_t, \boldsymbol{c}_{\text{sam}}, t, \theta_{\text{era}}), & \text{otherwise.} \end{cases}$$

and continues until $t = 0$, producing the final image $\boldsymbol{x}_0^*$. By localizing the erased concept via attention, masking its footprint, and overwriting it with noise, IRECE prevents concept reappearance while maintaining visual coherence. The approach operates entirely at inference time and readily extends to tasks such as object removal or replacement.

4 Experiments

4.1 Experimental Setup

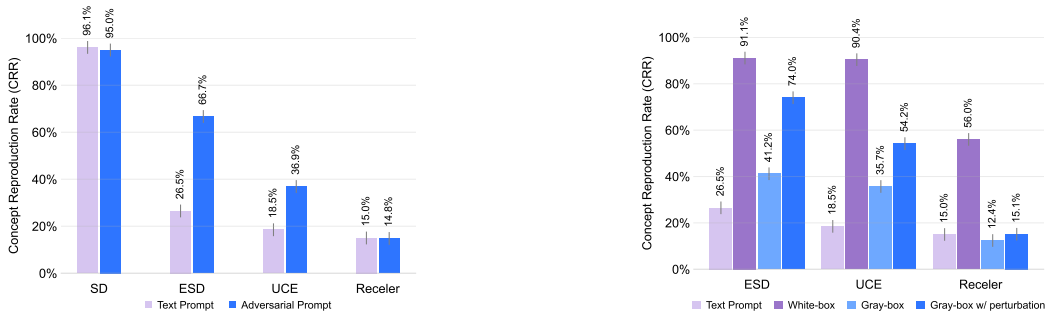
Dataset. We adopt Stable Diffusion (SD) v1.4 [2] as the base model and evaluate three representative erasure methods: ESD [19], UCE [20], and Receler [22]. Following prior work, we use CIFAR-10 class labels as target concepts. For evaluation, we construct four datasets.

1. **SD-Normal:** Constructed by generating images from SD using five prompt templates: "An image of `TARGET`", "A painting of `TARGET`", "A picture of `TARGET`", "A photo of `TARGET`", and simply "`TARGET`", each instantiated with 30 random seeds, yielding 150 prompts per class.
2. **SD-AdvPrompt:** Constructed by attacking the SD-Normal templates with Ring-A-Bell using the method's official configuration.
3. **SD-TI:** Constructed by learning a special embedding for each reference image in the SD-Normal set via Textual Inversion, optimized with a learning rate of 5e-4 for 1500 optimization steps, and then inserting the learned token into the same prompt templates.
4. **SD-LatentInv:** Constructed by applying DDIM inversion [25] to images in the SD-Normal dataset to obtain their initial latents.

Evaluation Metrics. We assess concept erasure using GroundingDINO [32] and report the *Concept Reproduction Rate (CRR)*, defined as the percentage of generated samples containing the erased concept. Lower CRR indicates stronger suppression.

4.2 Evaluation Results

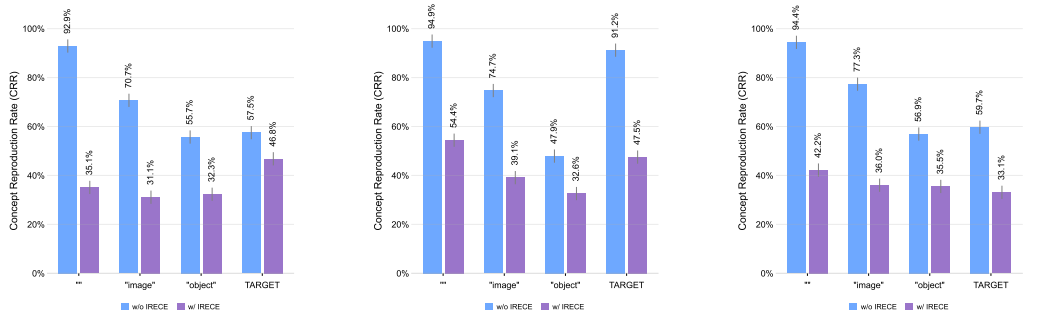
Text Prompt Evaluation. We evaluate concept erasure methods on the SD-Normal and SD-AdvPrompt datasets (Fig. 2a). Under standard text prompts, Stable Diffusion (SD) yields a high CRR of 96.1%, while ESD, UCE, and Receler reduce it to 26.5%, 18.5%, and 15.0%. However, robustness drops sharply under adversarial prompts: SD remains high at 95.0%, whereas ESD and UCE rise to 66.7% and 36.9%. Only Receler maintains a low CRR of 14.8%, confirming that adversarial training effectively mitigates such attacks.



(a) Text Prompt

(b) Learned Embedding

Figure 2: **Concept Reproduction Rate (CRR) for concept-erasure methods under two evaluation settings: (a) text prompts and (b) learned embeddings.** In (a), bar colors distinguish between original text prompts and adversarial prompts. In (b), bar colors denote white-box, black-box, and black-box with perturbation settings, with results from original text prompts included as a reference.



(a) ESD White-box Setting

(b) UCE with White-box Setting

(c) Receler with White-box Setting

(d) ESD with Gray-box Setting

(e) UCE with Gray-box Setting

(f) Receler with Gray-box Setting

Figure 3: **Comparison of Concept Reproduction Rate (CRR) with and without IRECE under latent-inversion evaluation.** Subfigures present results for three representative concept-erasure methods under white-box and gray-box settings. The horizontal axis groups correspond to different prompt types (“”, “image”, “object” and `TARGET`), and bar colors indicate whether IRECE is applied.

Learned Embedding Evaluation. We evaluate concept erasure methods on the SD-TI dataset (Fig. 2b). In the white-box setting, where embeddings are learned on the erased model, all methods show a sharp increase in CRR compared with the text-prompt case: ESD rises from 26.5% to 91.1%, UCE from 18.5% to 90.4%, and Receler from 15.0% to 56.0%, indicating that learned embeddings substantially weaken suppression. In the more realistic gray-box setting, where embeddings are trained on the standard model but tested on the erased one, CRR drops to 41.2% for ESD, 35.7% for UCE, and 12.4% for Receler. Notably, Receler falls below its text-prompt baseline, while ESD and UCE remain higher, suggesting persistent vulnerability. Adding small image perturbations (with a budget of 8) during embedding training further intensifies the attack, raising CRR to 74.0% for ESD, 54.2% for UCE, and 15.1% for Receler. Overall, learned embeddings and perturbations both compromise erasure robustness, with Receler remaining the most resilient.

Figure 4: **Comparison of erased models with the plug-and-play IRECE module across 10 target concepts.** Results are generated in the white-box setting with the unconditional prompt. The first column shows the reference images for latent inversion, and the remaining columns display outputs from different concept erasure methods.

Inverted Latent Evaluation. We evaluate concept erasure methods on the SD-LatentInv dataset (Fig. 3). In the white-box setting, latent inversion effectively bypasses erasure, with the unconditional prompt driving CRR above 92% for all methods. The "image" strategy also achieves over 70%, and even the `TARGET` strategy exceeds 50%, indicating that inversion substantially weakens suppression regardless of prompt type. In the gray-box setting, CRR remains higher than text-prompt baselines. The unconditional prompt continues to be the strongest, raising CRR to 57.3% for ESD, 95.2% for UCE, and 79.0% for Receler. These results show that latent inversion exposes the most severe vulnerability: by initializing the generative process from concept-containing latents, erased concepts frequently re-emerge, and even adversarially trained methods like Receler struggle to remain robust. This highlights the inherent limitation of current text-based defenses.

4.2.1 Inference-time Robustness Enhancement for Concept Erasure (IRECE)

Quantitative results in Fig. 3 show that IRECE consistently improves robustness across all settings. The largest gain occurs under the white-box latent inversion with unconditional prompt, where CRR drops by over 40% for all methods. These results demonstrate that IRECE restores robustness even under the most challenging attacks. Qualitative examples in Fig. 4 further confirm its effectiveness. For classes such as *airplane*, *bird*, *deer*, *ship*, and *truck*, the target concept is almost entirely removed, while for *automobile*, *cat*, *dog*, *frog*, and *horse*, it is replaced with alternative content. In both cases, non-target regions remain intact and transitions appear coherent, indicating that IRECE not only strengthens robustness but also enables controlled object removal and replacement.

5 Conclusion

We present a multimodal framework to evaluate the robustness of concept-erasure methods in diffusion models. While prior approaches perform well on text-prompt evaluations, they degrade sharply with learned embeddings and nearly fail under image latent inversion, revealing critical vulnerabilities. To address this, we propose IRECE, a plug-and-play inference-time module that restores robustness without retraining. Experiments show that IRECE reduces CRR by up to 40% under white-box latent inversion while preserving visual quality, establishing it as an effective defense against multimodal attacks. Beyond robustness, IRECE also supports targeted object removal and replacement, emphasizing the need for reliable concept erasure in generative models.

Acknowledgements

This research is supported by National Science and Technology Council, Taiwan (R.O.C) under the grant numbers NSTC-114-2634-F-001-001-MBK, NSTC-112-2222-E-001-001-MY2, 114-2221-E-002-182-MY3, and 113-2221-E-002-201 and Academia Sinica under the grant number of AS-CDA-110-M09 and AS-IAIA-114-M10.

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022.

[4] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint*, 2021.

[5] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2025.

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint*, 2022.

[8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023.

[10] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] Matt Blaszczyk, Geoffrey McGovern, and Karlyn Stanley. Artificial intelligence impacts on copyright law. *RAND*, 2024.

[14] Michael M Grynbaum and Ryan Mac. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 2023.

[15] Tatum Hunter. Ai porn is easy to make now. for women, that's a nightmare. *The Washington Post*, 2023.

[16] Laura Herijgers. Stable diffusion 3 is a step back for ai images of humans, 2024.

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

[18] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[19] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[20] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[21] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[22] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

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

[24] Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *International Conference on Learning Representations (ICLR)*, 2023.

[25] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[27] Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13393–13402, 2025.

[28] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Cheng Chiu. Prompting4debugging: red-teaming text-to-image diffusion models by finding problematic prompts. In *International Conference on Machine Learning (ICML)*, 2024.

[29] Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint*, 2023.

[30] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations (ICLR)*, 2024.

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.