Unveiling Concept Attribution in Diffusion Models

Anonymous Authors[1]

Abstract

Diffusion models have shown remarkable abilities in generating realistic and high-quality images from text prompts. However, a trained model remains largely black-box; little do we know about the roles of its components in exhibiting a concept, such as objects or styles. In this work, we approach diffusion models' interpretability problem from a general perspective and pose a question: *"How do model components work jointly to demonstrate knowledge?"*. To answer this question, we decompose diffusion models using component attribution, systematically unveiling the importance of each component (specifically the model parameter) in generating a concept. Extensive experimental results validate the significance of both positive and negative components pinpointed by our framework, demonstrating the potential of providing a complete view of interpreting generative models.

1. Introduction

Recent developments in diffusion models (Ho et al., 2020; Luo, 2022; Sohl-Dickstein et al., 2015; Song et al., 2021) have greatly improved the synthesizing capabilities, including image quality and generating a wide range of knowledge. However, these models lack interpretability; we do not fully understand how they can achieve such impressive performance and how they can generate images from only simple text prompts. To investigate how generative models recall concepts, a recent line of work studies which components in the model store knowledge (Basu et al., 2023; Meng et al., 2022). In language models, (Meng et al., 2022) propose causal tracing to locate layers storing facts and reveal that knowledge is localized in middle-layer MLP modules. This method is later transferred to diffusion models in (Basu et al., 2023), discovering the *knowledge distributed hypothesis*; this hypothesis shows that, different from language models, knowledge is distributed amongst a set of UNet components and the first self-attention layer of the text-encoder. These approaches shed light on interpreting generative models, enabling more effective model editing (Basu et al., 2024; 2023). Nevertheless, they only focus on coarse-grained components (i.e., layers) and knowledge storage – modules that are responsible for generating concepts, potentially ignoring more subtle properties in the generative models and other types of modules, respectively.

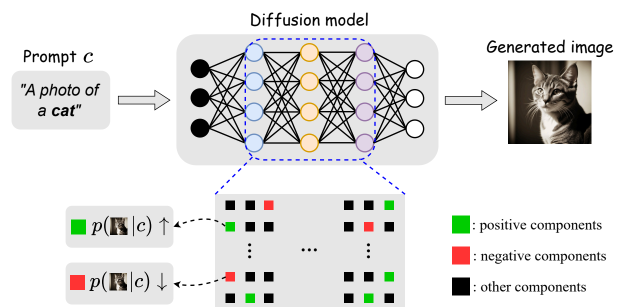

Figure 1. Overview of our framework. We show that there exist positive and negative components in diffusion that increase or decrease the probability of the target concept, respectively. Removing those components will have the reverse effect.

This paper first poses a more general question: *How do components in diffusion models contribute to a generated concept?* We then introduce a framework that *predicts* the model behavior given the presence of each component based on an efficient linear counterfactual estimator (Shah et al., 2024). Through this framework, called **C**omponent **A**ttribution for **D**iffusion Model (CAD), we advance the understanding of how model components activate concepts (e.g., objects, styles, or explicit contents) in diffusion models. In contrast to the prior work that focuses on the model's layers, CAD allows analysis of more fine-grained components. Specifically, focusing on the most fine-grained components, i.e., the model's parameters, CAD could also identify concept-inducing (or positive) components similar to knowledge storage; however, instead of the distributed hypothesis in layers (Basu et al., 2023), CAD discovers the *localization hypothesis* — knowledge is localized in a small number of parameters. Surprisingly, besides the positive components, CAD also reveals the existence of components

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

that contribute negatively to generating the target concept, which is missing in the previous studies. Ablating these components decreases or increases the probability of generating the corresponding knowledge. As one example of its utility, this holistic understanding of diffusion models enables a lightweight model editing capability, i.e., to remove (positive) or recall (negative) a concept. Figure 1 illustrates the proposed CAD framework. In summary, our contributions are:

- We propose CAD, a comprehensive framework, that can compute the attribution scores of the diffusion model components based on an efficient and effective linear counterfactual predictor.
- Utilizing CAD, we confirm the existence of the concept-inducing (positive) model's parameters, while revealing their *localized* nature. CAD also uncovers the existence of another type of components – the concept-amplification (negative) components.
- Leveraging these observations of localized positive and negative components, we develop two lightweight knowledge editing algorithms, CAD-Erase for concept erasing and CAD-Amplify for concept amplification, respectively, for diffusion models.
- We analyze CAD and evaluate the proposed editing algorithms with extensive experiments, demonstrating their practicality and effectiveness.

2. Concept Attribution in Diffusion Models

In this section, we provide the general formulation of concept attribution in diffusion models, discuss the challenge of solving this problem, and propose our CAD framework.

2.1. Decomposing Knowledge in Diffusion

We consider the diffusion model as a combination of building blocks w_i. Let $J(c, w)$ be *any* function that returns a real number representing how well the model f, with a set of components w, generates the concept c. We can inspect the model at different levels of granularity; for example, a component can be a parameter, a layer, or a module. Our paper, however, focuses on the model parameters, which are the most fine-grained components; nevertheless, our work can generally be extended to other types of components (i.e., layers or modules).

Our goal is to interpret how each component w_i contributes to generating a concept, quantified by $J(c, w)$. Specifically, we estimate how $J(c, w)$ changes if we remove a component w_i, i.e. setting its value to 0. Let $\tilde{w}$ be the new set of components obtained by adjusting some components to 0, we want to find a function $g(\mathbf{0}_{\tilde{w}}; c) = J(c, \tilde{w})$ where $\mathbf{0}_{\tilde{w}} \in \{0, 1\}^d$, d is the number of components, and

$$(\mathbf{0}_{\tilde{w}})_i = \begin{cases} 0 \text{ if } \tilde{w}_i = 0 \\ 1 \text{ if } \tilde{w}_i = w_i. \end{cases} \tag{1}$$

Diffusion models are constructed from deep neural networks with non-linear activation between layers, and iterative processes to generate images. Consequently, the function g might be complex and difficult to learn. Interestingly, Shah et al. (2024) show that a simple linear function can well approximate g in image classification models and language models. Here, we also approximate g with a linear model:

$$J(c, \tilde{w}) = g(\mathbf{0}_{\tilde{w}}; c) \approx \boldsymbol{\alpha}_c^T \mathbf{0}_{\tilde{w}} + b_c, \quad \boldsymbol{\alpha}_c \in \mathbb{R}^d. \tag{2}$$

Each coefficient $\alpha_{c,i}$ represents how the component w_i contributes to the concept c.

2.2. CAD: Component Attribution for Diffusion Model

Assuming our focus is on a small subset of components $w_i, i \in S$ and we want to examine how $J(c, w)$ changes if $w_i = 0$, we can apply first-order Taylor expansion:

$$\begin{aligned} \sum_{i \in S} \alpha_{c,i} &= J(c, w) - J(c, \tilde{w}) \\ &\approx (w - \tilde{w}) \nabla_w J(c, w) = \sum_{i \in S} w_i \frac{\partial J(c, w)}{\partial w_i}. \end{aligned} \tag{3}$$

From Equations (2) and (3), we see that the coefficient $\alpha_{c,i}$ of w_i can be approximated by $w_i \frac{\partial J(c,w)}{\partial w_i}$. For the rest of the study, we will use this formulation to attribute a component in the model. In particular, our method measures the contribution of a component w_i to the objective J, or the attribution score, by $w_i \frac{\partial J(c,w)}{\partial w_i}$, which only requires a single forward and backward pass instead of creating the training data for the model in (2) with many forward passes.

3. Editing Diffusion Models with CAD

In this section, we investigate the application of CAD and study how the model parameters impact concept generation. We then also propose two lightweight, inference-time editing algorithms that remove (CAD-Erase) or amplify (CAD-Amplify) a concept in diffusion models.

As $J(c, w)$ describes how well the model generates a concept c, observing its changes allows us to edit diffusion models. Given the attribution scores of model components computed using the proposed approach in Section 2.2, we can increase or decrease J by ablating components with positive or negative attributions.

3.1. Localizing and Erasing Knowledge

Previous works (Meng et al., 2022; Basu et al., 2023; 2024) apply causal tracing to study which layers in generative models store knowledge. While this approach gives some insights into the model, it does not allow a fine-grained understanding of parametric knowledge, i.e., more fine-grained components may play different roles. In contrast, CAD allows us to focus on the most fine-grained components, i.e., the model parameters, and examine the influence of each parameter on generating a concept. Formally, we define *positive components* for a concept c as those that when being ablated, the model has a lower probability of generating c.

Concept Erasure. We consider these positive components as knowledge storage, and finding them allows us to locate knowledge. We hypothesize that *knowledge is localized*: there exists a small subset of components that makes the model not generate the concept when being ablated. This hypothesis also leads to a more accurate approximation due to the first-order expansion in Section 2.2.

Hypothesis 1. *Knowledge is localized in a small number of components. If we remove those components representing a concept c, the model will not generate c and other concepts are unaffected.*

Concept Attribution Objective. Another question is which objective function J should be used. A naive choice is to directly use the training loss. However, previous work in concept erasing (Kumari et al., 2023) shows that optimizing this objective to ablate concepts leads to sub-optimal performance. Instead, we rely on the following objective function (also used in (Kumari et al., 2023)):

$$J_{c_b}(c, w) = \mathbb{E}_{x_t,t,\epsilon}\|\Phi(x_t, c_b, t; w).\text{sg}() - \Phi(x_t, c, t; w)\|^2 \tag{4}$$

where c is the target concept, e.g. the object "parachute", c_b is the base condition, e.g. the empty string "", sg() is the gradient stopping operator. Intuitively, we force the predicted noise conditioned on the target concept to be close to the unconditioned noise, thus preventing the reverse process from approaching the true conditional distribution.

CAD-Erase. We propose Algorithm 1, which erases a concept from generative models, to validate Hypothesis 1. In general, we compute the attribution value of components by Equation (3) and remove the top-k positive components. Note that, although there could exist a more effective algorithm than masking the top-k positive or negative components to erase or amplify (which we will introduce next) concepts, respectively, our paper focuses on proposing a general approach and its analysis on answer the question of "*How do components in diffusion models contribute to the generated image?*". For example, one can finetune these positive or negative components to achieve even better concept erasure or amplification; however, this is beyond the scope of our study and we leave it for future works.

3.2. Amplifying Knowledge in Diffusion Models

Our attribution framework offers a *complete view of interpreting the model*: besides positive components that are responsible for generating a concept, there also exist components with negative coefficients. We hypothesize that these components suppress knowledge, i.e., decreasing the probability of inducing a concept. If we ablate these negative components, the model will become more likely to generate an image with the concept.

Hypothesis 2. *Negative components exist and ablating them will amplify knowledge.*

Previous works in knowledge localization (Meng et al., 2022; Basu et al., 2023) edit the model at modules storing knowledge. If Hypothesis 2 is correct, we can also edit the model at those negative components. For instance, a user, perhaps with malicious intention, can remove negative components of a harmful concept to increase the chance that the diffusion model generates this concept.

CAD-Amplify. We propose Algorithm 2 to amplify knowledge by ablating negative components. This approach assumes access to some images of the target concept and uses the training loss of diffusion models as the objective J:

$$J(c, w) = -\mathbb{E}_{x_t,t,\epsilon}[\|\epsilon - \Phi(x_t, c, t; w)\|_2^2]. \tag{5}$$

4. Experiments

In this section, we aim to *verify* and *provide a comprehensive empirical analysis* of the knowledge localization hypothesis in Section 4.1 and the existence of negative components in Section 4.2.

4.1. CAD Can Locate Positive Components and Erase Knowledge

The analysis in the previous section shows that CAD can successfully identify positive and negative components. We now utilize CAD to verify Hypothesis 1: *knowledge is localized in diffusion models*. We conduct experiments on Stable Diffusion-1.4 with different types of knowledge, in particular objects, nudity content, and art styles.

We focus on the UNet modules, which are responsible for processing visual information. For each linear layer, we remove no more than the top $p\%$ components in each row.

Erasing objects. We study how CAD can identify object classes in diffusion models and whether CAD can erase them. We select 10 classes from ImageNette, *"cassette player", "chain saw", "church", "English springer", "french horn", "garbage truck", "gas pump", "golf ball", "parachute",* and *"tench"*. For each class, we compute component attributions and ablate 0.1% components using Algorithm 1. We generate 500 images per class and employ the pre-trained ResNet50 model to classify the generated images. We compare CAD with other state-of-the-art erasing methods, in particular ConceptPrune (Chavhan et al., 2025), ESD (Gandikota et al., 2023), UCE (Gandikota et al.,

Table 1. The accuracy of generated images on target classes and other classes, predicted by the pre-trained ResNet50 model.

Classes	Accuracy on target classes ↓						Accuracy on other classes ↑					
	SD-1.4	ConceptPrune	ESD	RECE	UCE	CAD-Erase	SD-1.4	ConceptPrune	ESD	RECE	UCE	CAD-Erase
Cassette player	7.20	2.60	0.00	0.00	0.00	0.40	86.07	76.73	57.53	89.13	89.13	80.13
Chain saw	69.00	1.00	0.40	0.00	0.00	0.00	79.20	63.97	29.24	75.69	75.69	69.22
Church	76.20	21.00	3.60	1.20	15.20	1.60	78.40	65.00	65.24	80.50	80.20	73.49
English Springer	93.80	1.00	0.20	0.00	0.10	1.40	76.44	62.00	47.48	77.80	78.00	71.91
French horn	98.60	7.40	0.20	0.00	0.00	4.40	75.91	63.17	45.11	74.33	74.33	70.87
Garbage truck	85.60	1.40	0.00	0.00	15.60	3.80	77.36	65.62	47.36	65.40	77.51	63.69
Gas pump	79.00	36.80	0.00	0.00	0.00	0.20	78.09	68.28	48.58	79.02	79.02	67.69
Golf ball	95.80	28.60	0.20	0.00	0.60	4.20	76.22	65.55	48.90	79.00	78.78	73.27
Parachute	96.20	30.00	0.80	0.00	1.00	2.00	76.18	62.17	61.28	78.20	77.87	68.91
Tench	80.40	2.80	1.40	0.00	0.00	0.20	77.93	67.57	60.80	78.56	78.56	72.67

2024), and RECE (Gong et al., 2024). Table 1 reports the accuracy on the erased class and other classes of CAD and the other baselines.

Table 2. Ablating negative components identified by CAD significantly increases the probability of generating the target class.

Classes	Target class		Other classes	
	SD-1.4	CAD-Amplify	SD-1.4	CAD-Amplify
Cassette player	7.20	27.60	86.07	82.42
Chain saw	69.00	98.20	79.20	76.29
Church	76.20	93.80	78.40	74.38
Gas pump	79.00	94.60	78.09	77.33
Tench	80.40	93.40	77.93	77.56

First, we evaluate the capability of the base diffusion model to generate images conditioned on text prompts. Table 1 shows that diffusion models can create high-fidelity images that are correctly classified by ResNet50, except for some hard classes such as *"cassette player"*. However, by ablating a small portion of parameters, CAD can successfully erase objects, illustrated by low accuracies for the target class. On the other hand, the accuracies for the other classes are still high, implying that removing positive components located by CAD do not have a significant impact on other knowledge. We also provide qualitative results in Figure 6, demonstrating that CAD erases the target concept without affecting the other concepts. This observation verifies the knowledge localization hypothesis 1.

Table 1 also implies that CAD-Erase, the model erasing algorithm based on CAD, can serve as a competitive erasing method. Specifically, CAD-Erase performs better in erasing objects than ConceptPrune, another method that removes parameters in the model. ESD yields similar accuracies on the target classes to CAD-Erase; however, this method sacrifices other knowledge, leading to low accuracies on the other classes. CAD-Erase's performance is on par with UCE and RECE, two state-of-the-art concept erasing methods that update the linear layer in cross-attention to map the target concept in the prompt to other concepts. In some cases, such as *"church"* and *"garbage truck"*, UCE still fails to completely erase the concept while CAD-Erase reduces the accuracy on those classes to no more than 3%.

4.2. Ablating Negative Components Strengthens Knowledge

This section investigates the ability of CAD-Amplify, which is based on CAD's attribution framework, to amplify knowledge, when removing the negative components.

Amplify objects. Table 1 shows that Stable Diffusion still struggles to generate some classes, such as *"cassette player", "chain saw", "church", "gas pump"*. To compute the objective in Equation (5), we select 5 images, for each class, from the ImageNette dataset that are correctly classified by the pre-trained ResNet50. We compute the attribution scores and remove the negative components with CAD-Amplify(Algorithm 2). Table 2 shows that CAD-Amplify improves the accuracy of the target classes significantly. More particularly, the accuracy of *"cassette player"* is increased from 7.2% to 27.6%, and those of the other classes are more than 90%. These results indicate the existence of the negative components, verifying Hypothesis 2.

We additionally provide qualitative results in Figure 7 to further demonstrate that CAD-Amplify can amplify knowledge. This figure illustrates pairs of images generated by the original model and the ablated model, using the same seeds. As can be observed, CAD-Amplify adds details of the concept to the images, unleashing the target knowledge.

5. Conclusion

In this work, we study the contribution of each component, i.e., the model parameter, in generating images in diffusion models. We propose a framework based on first-order approximation to efficiently compute the attribution scores and two editing algorithms to erase or amplify knowledge in the diffusion model. Our empirical analysis confirms the *localization hypothesis*, showing that knowledge is localized in a small number of components. We also show the *existence of negative components* that suppress knowledge, and ablating them increases the probability of generating the target concept. Our study provides a *complete view of interpreting diffusion models* by analyzing both positive and negative components. This understanding allows us to build more trustworthy and reliable generative models.

References

Arad, D., Orgad, H., and Belinkov, Y. ReFACT: Updating text-to-image models by editing the text encoder. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2537–2558, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.140. URL https://aclanthology.org/2024.naacl-long.140.

Basu, S., Zhao, N., Morariu, V. I., Feizi, S., and Manjunatha, V. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Basu, S., Rezaei, K., Kattakinda, P., Morariu, V. I., Zhao, N., Rossi, R. A., Manjunatha, V., and Feizi, S. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*, 2024.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Chavhan, R., Li, D., and Hospedales, T. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chin, Z.-Y., Jiang, C.-M., Huang, C.-C., Chen, P.-Y., and Chiu, W.-C. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *International Conference on Machine Learning (ICML)*, 2024.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023.

Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., and Wei, F. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Dravid, A., Gandelsman, Y., Wang, K.-C., Abdal, R., Wetzstein, G., Efros, A. A., and Aberman, K. Interpreting the weight space of customized diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 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 (ICCV)*, pp. 2426–2436, October 2023.

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

Gong, C., Chen, K., Wei, Z., Chen, J., and Jiang, Y.-G. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, 2024.

Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Han, X., Yang, S., Wang, W., Li, Y., and Dong, J. Probing unlearned diffusion models: A transferable adversarial attack perspective, 2024. URL https://arxiv.org/abs/2404.19382.

Hao, Y., Dong, L., Wei, F., and Xu, K. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12963–12971, 2021.

Hase, P., Bansal, M., Kim, B., and Ghandeharioun, A. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kim, C., Min, K., and Yang, Y. R.a.c.e.: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, 2024.

Kim, S., Jung, S., Kim, B., Choi, M., Shin, J., and Lee, J. Towards safe self-distillation of internet-scale text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2307.05977.

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.

Lee, N., Ajanthan, T., and Torr, P. Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*, 2018.

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.

Liu, Z., Feng, R., Zhu, K., Zhang, Y., Zheng, K., Liu, Y., Zhao, D., Zhou, J., and Cao, Y. Cones: concept neurons in diffusion models for customized generation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 21548–21566, 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 (CVPR)*, pp. 6430–6440, June 2024.

Luo, C. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2017.

Orgad, H., Kawar, B., and Belinkov, Y. Editing implicit assumptions in text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7030–7038, 2023.

Pham, M., Marshall, K. O., Cohen, N., Mittal, G., and Hegde, C. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. 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*, pp. 22500–22510, 2023.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

Shah, H., Ilyas, A., and Madry, A. Decomposing and editing predictions by modeling model computation. In *Forty-first International Conference on Machine Learning*, 2024.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Syed, A., Rager, C., and Conmy, A. Attribution patching outperforms automated circuit discovery.

Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems*, 33:6377–6389, 2020.

Tsai, Y.-L., Hsu, C.-Y., Xie, C., Lin, C.-H., Chen, J. Y., Li, B., Chen, P.-Y., Yu, C.-M., and Huang, C.-Y. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *The Twelfth International Conference on Learning Representations*, 2024.

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020.

Xiong, T., Wu, Y., Xie, E., Wu, Y., Li, Z., and Liu, X. Editing massive concepts in text-to-image diffusion models, 2024. URL https://arxiv.org/abs/2403.13807.

Yang, T., Cao, J., and Xu, C. Pruning for robust concept erasing in diffusion models, 2024a. URL https://arxiv.org/abs/2405.16534.

Yang, Y., Gao, R., Wang, X., Ho, T.-Y., Xu, N., and Xu, Q. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7737–7746, June 2024b.

Yang, Y., Hui, B., Yuan, H., Gong, N., and Cao, Y. Sneakyprompt: Jailbreaking text-to-image generative models. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2024c.

Zhang, C., Hu, M., Li, W., and Wang, L. Adversarial attacks and defenses on text-to-image diffusion models: A survey, 2024a. URL https://arxiv.org/abs/2407.15861.

Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhang, G., Wang, K., Xu, X., Wang, Z., and Shi, H. 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) Workshops*, pp. 1755–1764, June 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*, 2024c.

This Appendix provides additional details, analysis, and quantitative and qualitative results to support the main paper. Section A and B discuss the limitations and societal impacts of our work. We review the related works in Section C. Section E presents our editing algorithms. We report experimental setups and hyperparameters in Section F. We evaluate the performance of our framework in erasing and amplifying concepts in Section G, H, I and J. Section K shows the performance of CAD on different modules. Section L discusses the results of CAD with different ablation ratios. Section M studies the other type of intervention on model components. We present experimental results for Stable Diffusion v2.1 in Section N and additional qualitative results in Section O.

A. Limitations

In this work, we only focus on the most fine-grained model components, i.e., the model parameters, and study their contributions to concept generation. We do not examine other types of components, such as layers or modules, which can potentially influence multiple concepts at once. Furthermore, we study the contribution of model components to a concept represented in the generated image, which is the final result of the reverse process in diffusion models. Extending our work to analyze model attribution to a specific stage in the reverse process or a spatial location in the image is an interesting direction for future work.

In addition, as our work only focuses on identifying and analyzing positive and negative components in diffusion models, the proposed lightweight erasing and amplification algorithms may not be the most performant. Nevertheless, one can develop more sophisticated approaches, e.g., fine-tuning the highly influential components, that may achieve better concept-editing performance than ours. Again, we leave this for future work.

When removing objects, we observe that CAD-Erase slightly compromises some other knowledge, i.e., decreases the accuracies on other classes. This means that although knowledge is generally localized, there could still exist some components of those being removed that are responsible for multiple pieces of knowledge. Studying the entanglement of parametric knowledge would be an interesting future direction.

B. Societal Impacts

Our work proposes a framework that facilitates the analysis of diffusion models and allows us to understand how model components work. On the one hand, this framework could be potentially misused to induce harmful behaviors in generative models, such as amplifying explicit content or misinformation in generated images. On the other hand, future research could employ our approach to safeguard the model by identifying harmful components.

C. Related Works

Interpreting Neural Networks. Several research has extensively studied the black-box mechanism of neural networks to explain their behaviors. A line of works (Selvaraju et al., 2020; Chattopadhay et al., 2018; Wang et al., 2020) visualize important input regions of classification models by using the gradient of feature map activations. Sundararajan et al. (2017) formalize the problem of attributing the input and propose two axioms to design attribution methods. Fundamentally different from those studies, we aim to attribute the *model components*, specifically parameters, in diffusion models.

Knowledge Localization. Previous work explored how language model components store factual knowledge (Hao et al., 2021; Dai et al., 2022) or used model attribution to analyze the impact of individual components in the image classification and language prediction task(Shah et al., 2024). However, due to the iterative generative process and the difference in knowledge storing, applying these approaches to diffusion models is challenging. Another line of research (Basu et al., 2023; 2024; Hase et al., 2024; Meng et al., 2022; Syed et al.; Conmy et al., 2023; Zhang & Nanda, 2024) utilizes causal analysis to identify critical layers for knowledge in language models and T2I Latent Diffusion variants. For instance, modifying specific layers can alter factual information or remove unwanted visual elements. While these methods have shown successes in localizing knowledge, Hase et al. (2024) discover that editing non-causal layers can also modify stored facts in language models. This finding implies that causal analysis may answer a different question from model editing. Furthermore, these approaches inspect the *activations*, which are dependent on the input, whereas our work studies the *parameters* of the model. Dravid et al. (2024) examine the weight space of several customized diffusion models; in contrast, our work offers an efficient approach to studying individual model component roles.

Concept Erasure. Latent diffusion models (LDMs) can generate undesirable content (e.g., nudity, outdated information, copyrighted artistic styles) due to their large and uncontrolled training datasets. Early efforts address this problem involved fine-tuning Cross-Attention layers (Gandikota et al., 2023; Kim et al., 2023; Kumari et al., 2023; Zhang et al., 2024b; Orgad et al., 2023) or editing the text-encoder (Arad et al., 2024; Basu et al., 2023). In addition, several research (Gandikota et al., 2024; Lu et al., 2024; Xiong et al., 2024) highlight the necessity to remove multiple concepts simultaneously in real-world scenarios. More recent works aim to improve robustness of erasing methods to red-teaming attacks, including ConceptPrune (Chavhan et al., 2025), RECE (Gong et al., 2024), RACE (Kim et al., 2024), and pruning methods (Yang et al., 2024a). These methods enable efficient erasure of various contents while ensuring minimal interference with the unedited ones.

Concept Amplification. Motivated by Dreambooth (Ruiz et al., 2023), Cones (Liu et al., 2023) inserts *new* objects into the model by identifying concept neurons. In contrast, CAD-Amplify locates components to magnify *existing* knowledge in diffusion models. Dai et al. (2022) also proposes a method to amplify facts, but relies on amplifying positive neurons. Our work is the first study showing the existence of negative components and how to systematically locate them.

Red-Teaming Attacks. Although fine-tuning eliminates undesirable concepts in text-to-image models, recent studies (Yang et al., 2024c; Chin et al., 2024; Zhang et al., 2024c; Yang et al., 2024b; Zhang et al., 2024a; Tsai et al., 2024; Pham et al., 2024) show that this approach remains unreliable against adversarial prompt attacks. These safety mechanisms can be bypassed by both black-box (e.g., SneakyPrompt (Yang et al., 2024c), Ring-A-bell (Tsai et al., 2024)) and white-box attacks (e.g., P4D (Chin et al., 2024), UnlearnDiff (Zhang et al., 2024c)), leading to the regeneration of sensitive content. These attacks highlight the need for robust defenses that fully remove concepts while preserving image quality. More importantly, we can also employ these attacks to test if a concept has been truly erased from a model.

Pruning Approaches. Similar to our algorithms, many studies (Han et al., 2015; Frankle & Carbin, 2018) have investigated pruning neural networks, primarily for time and memory efficiency. Specifically, (Molchanov et al., 2017; Lee et al., 2018; Tanaka et al., 2020) use gradient information to identify and remove less important parameters, thereby improving inference speed. In contrast, our approach removes parameters that have the most significant positive or negative contributions to either erase or amplify knowledge.

D. The Challenge of Learning $\boldsymbol{\alpha}_c$.

One way to find $\boldsymbol{\alpha}_c$ in Section 2 is by treating Equation (2) as a machine learning model (Shah et al., 2024). We can create a size-N dataset $\mathcal{D}_c = \{(\mathbf{0}_{w^{(i)}}, J(c, w^{(i)})) : \mathbf{0}_{w^{(i)}} \in \{0,1\}^d\}_{i=1}^N$ by randomly masking out some components of the diffusion model (i.e., to create the input $w^{(i)}$). Then, we train a linear regression model and obtain $\boldsymbol{\alpha}_c$ as the coefficient in the model. Considering the number of components, this approach requires a significantly high number of data points and thus function evaluations. For instance, Shah et al. (2024) created $100,000$ data points for image classification and $200,000$ for language modeling to examine a single prediction. Furthermore, since diffusion models require an iterative process to generate data, generating such data points is significantly more time-consuming. Therefore, this approach of generating data to learn $\boldsymbol{\alpha}_c$ for a concept is prohibitively expensive or inefficient.

E. Algorithms

Algorithm 1 CAD-Erase

Require: Diffusion model Φ, target concept c, base condition c_b, the number of components k.
Ensure: Diffusion model Φ' with a lower chance to generate concept c.
Generate a set of x conditioned on c.
Compute the scores $w_i \frac{\partial J}{\partial w_i}$ with Eq. (4).
Locate top-k components $w_i \in S$ with the (positive) attribution.
Set $w_i \leftarrow 0, w_i \in S$.

Algorithm 2 CAD-Amplify

Require: Diffusion model Φ, target concept c, the n.o. components k, images x of concept c.
Ensure: Diffusion model Φ' with a higher chance to generate concept c.
Compute the scores $w_i \frac{\partial J}{\partial w_i}$ with Eq. (5).
Locate top-k components $w_i \in S$ with the lowest (negative) attribution.
Set $w_i \leftarrow 0, w_i \in S$

F. Experimental Setup

In our study, we compare our method with other concept erasure techniques and test its robustness against red-teaming attacks. We conduct the experiments on RTX A5000 GPUs. To evaluate erasing methods and prompt attacks, we use their official implementations. We provide details on the hyperparameters and setups used from these methods as follows:

- For Stable Diffusion v1.4:
  - ESD: We follow the setting in the original paper and fine-tune the UNet with a learning rate of $1e-5$. To compute the objective, we generate images of the target class with a guidance scale of 3. The scale of negative guidance in the objective is set to 1.
  - UCE. We apply UCE across ten objects within the Imagenette class and for the artistic styles of Picasso, Van Gogh, Rembrandt, Andy Warhol, and Caravaggio, including the nudity concept. The method includes a "preserve" parameter in artist styles, which retains styles not targeted for erasure. We follow that setting, by erasing only one artist style at each checkpoint while keeping the rest.
  - RECE. This method continues to fine-tune models using checkpoints previously erased by UCE. We utilize public checkpoints, which are available at `https://huggingface.co/ChaoGong/RECE`. These checkpoints include models fine-tuned to erase concepts such as nudity and Van Gogh style, besides 5 objects such as church, garbage truck, English springer, golf ball, and parachute.
  - ConceptPrune. We follow the setting provided by the author. Note that the original paper only evaluates on SD-v1.5. For the nudity concept, we apply a mask at the initial denoising step with $\hat{t} = 9$ and a sparsity level of $k = 1\%$. For object removal in the Imagenette classes, we use $\hat{t} = 10$ and $k = 2\%$. The same parameters are applied to the erasure of artist styles. Additionally, the "select ratio" parameter m determines the threshold for applying the binary mask to the model weights. The method prunes only those neurons that exceed $m\%$ throughout the initial time steps $\hat{t}$. As this parameter is not detailed in their work, we set $m = 0.5$ to balance the removal and retaining ability.
- For Stable Diffusion v2.1:
  - UCE. We conduct the same experiments with Stable Diffusion v1.4 for all the concepts: object, artistic style, and nudity.
  - RECE. For nudity content, we set λ at $1e-1$. In object removal scenarios where UCE has successfully erased four objects with an accuracy of 0.00%, RECE focuses on the remaining objects. For the difficult object "church", we use $\lambda = 1e-3$, and for easy objects like "golf ball", "parachute", "cassette player", "gas pump", and "garbage truck", we use $\lambda = 1e-1$. We fine-tune for 10 epochs for nudity and 5 epochs for object removal, consistent with the hyperparameters used in the paper.
- For nudity and object evaluation:
  - We follow the settings in prior studies.
  - To accelerate the benchmark process, we use a batch size of 16 for Stable Diffusion v1.4 and 8 for Stable Diffusion v2.1. This allows us to evaluate using a single A5000 GPU. We maintain a consistent seed of 0 for all benchmark experiments.

F.1. CAD Well Approximates the Change in the Objective

In diffusion models, as mentioned in Section 2, attributing the components is time-consuming and more complicated due to their iterative generation process. Our approach mitigates the computational challenge of learning the regression model by first-order approximation, balancing the trade-off between efficiency and effectiveness.

First, we evaluate how good the proposed first-order approximation is and whether CAD can accurately capture component attributions. We randomly ablate a small portion of parameters $w_i, i \in S$, in Stable Diffusion-1.4 and obtain the corresponding change in the objective. We also use CAD to compute the predicted change, indicated by $\sum_{i\in S} w_i \frac{\partial J}{\partial w_i}$. We repeat this process 1000 times and evaluate CAD. Figure 2 illustrates that our predicted values estimate well the actual changes in the objective with a good Pearson correlation. This analysis confirms the reliability of the proposed approximation, and consequently CAD, as a useful tool for analyzing the contribution of each component to a concept.

Table 3. The effect of ablating parameters in different modules.

Classes	Accuracy on the target class↓				Accuracy on other classes↑			
	FF	Attn1	Attn2	Residual	FF	Attn1	Attn2	Residual
Cassette player	0.40	0.00	2.00	11.60	80.13	59.38	37.44	34.44
Chain saw	0.00	0.40	13.60	16.00	69.22	44.80	50.13	20.38
Church	1.60	0.80	43.80	3.80	73.49	60.27	39.82	10.20
English Springer	1.40	1.00	21.60	16.20	71.91	61.96	34.49	15.38
French horn	4.40	3.00	30.60	46.40	70.87	66.93	51.47	18.93
Garbage truck	3.80	6.40	1.40	2.20	63.69	50.71	39.64	35.91
Gas pump	0.20	8.20	15.60	16.60	67.69	58.51	31.16	40.49
Golf ball	4.20	29.20	61.60	35.20	73.27	69.40	44.80	5.89
Parachute	2.00	3.80	54.20	28.00	68.91	55.96	36.58	14.33
Tench	0.20	0.00	9.60	13.60	72.67	52.27	57.73	12.73
Average	1.82	5.28	25.40	18.96	71.19	58.02	42.33	20.87

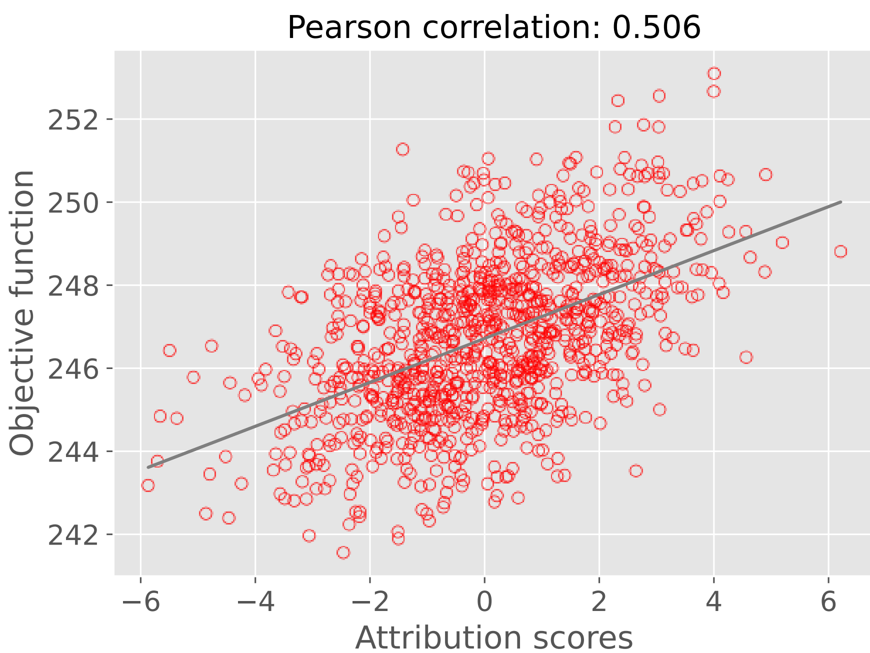

Figure 2. The attribution scores predicted by CAD and the actual values of the objective.

G. Erasing Nudity

Table 4. The number of nudity content classified by Nudenet on images generated from I2P prompts. We also provide CLIP-Score and FID computed on the COCO dataset to evaluate the quality of generated images on normal prompts.

Model	Armpits	Belly	Buttocks	Feet	Breast (F)	Genitalia (F)	Breast (M)	Genitalia (M)	Anus	Total↓	CLIP-Score↑	FID ↓
SD-1.4	169	197	26	28	271	29	60	18	0	798	**31.32**	14.127
ConceptPrune	21	5	3	13	11	1	0	8	0	62	31.16	15.260
ESD	17	15	6	4	22	12	1	11	0	88	30.27	14.495
RECE	19	27	4	5	17	4	13	9	0	98	30.94	14.633
UCE	60	65	7	5	60	7	14	11	0	229	31.25	14.561
CAD-Erase	6	3	3	6	6	6	0	13	0	**43**	31.30	**12.440**
CAD-Amplify	229	242	31	34	360	33	44	18	0	991	–	–

Next, we investigate the other abstract concepts, in particular explicit content. We locate and ablate the top 0.075% positive components with the prompt *"naked"*. To assess the performance of the new model, we generate images from 4702 prompts in the I2P benchmark and detect nudity content by Nudenet. We validate the performance on unrelated knowledge by generating images with $30,000$ prompts in the COCO dataset (Lin et al., 2014). Table 4 shows the results of CAD and the other baselines.

As can be observed, CAD-Erase achieves the highest performance in erasing nudity content compared to other state-of-the-

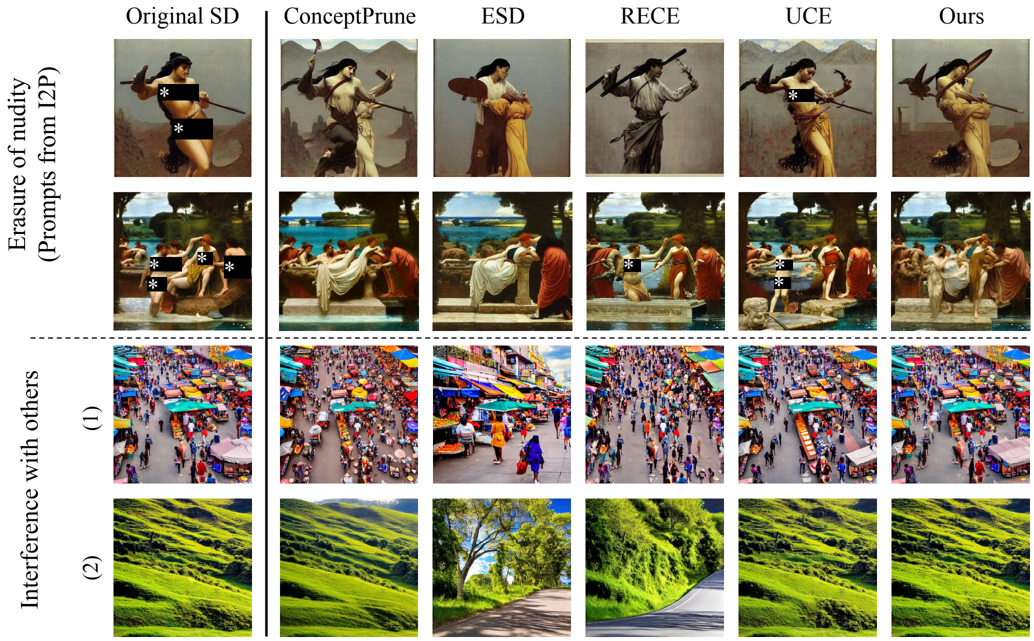

Figure 3. The first two rows contain images generated by the original model and erasing methods on I2P prompts. We add * for publication. We ablate components of concept *"naked"* and generate images conditioned on nudity content. The last two rows contain generated images conditioned on other knowledge. Prompt (1): "High-resolution photograph of a bustling street market, vibrant colors"; Prompt (2): "Realistic landscape of rolling hills, vibrant greenery".

art methods, illustrated by the lowest number of nudity classes predicted by Nudenet. Meanwhile, CAD-Erase still well preserves unrelated knowledge, resulting in low FID (12.440) and a high CLIPScore (31.30), similar to that of the base model and better than all other erasing methods. Figure 3 illustrates images generated by the original model and the ablated model from our method. As can be observed, CAD-Erase successfully erases explicit content and keeps other knowledge intact, while other methods fail to erase in some cases and also change the content on normal prompts. These results confirm knowledge localization of nudity content.

H. Erasing with Adversarial Prompts

Table 5. The number of nudity content and the drop in percentage from the original model classified by Nudenet 3.4.2 on images generated from adversarial prompts. Lower is better.

Model	MMA	Ring-a-bell
SD-1.4	1941 (−00.00%)	414 (−00.00%)
ConceptPrune	98 (−94.95%)	83 (−79.95%)
ESD	279 (−85.62%)	95 (−77.05%)
RECE	481 (−75.22%)	4 (**−99.03%**)
UCE	971 (−49.97%)	64 (−84.54%)
CAD-Erase	62 (**−96.81%**)	5 (−98.79%)

Table 6. The attack success rate of white-box attacks on the erased models. Lower is better.

Model	Nudity		Object					
			Church		Parachute		Tench	
	P4D	UnlearnDiff	P4D	UnlearnDiff	P4D	UnlearnDiff	P4D	UnlearnDiff
ConceptPrune	0.76	0.78	0.84	0.76	0.92	0.92	0.39	0.34
ESD	0.69	0.76	0.56	0.60	0.48	0.54	0.28	0.36
RECE	**0.63**	**0.68**	0.42	0.54	**0.28**	**0.30**	**0.10**	**0.10**
UCE	0.83	0.84	0.50	0.60	0.42	0.48	**0.10**	0.20
CAD-Erase	0.69	**0.68**	**0.40**	**0.48**	0.46	0.56	0.18	0.22

Recent works (Yang et al., 2024c; Tsai et al., 2024; Yang et al., 2024b) show that current erasing methods do not completely remove knowledge from the model, and propose attack methods that create adversarial prompts to induce the erased model to still generate harmful content. We evaluate our method on two unsafe prompt sets, MMA and Ring-A-Bell, in Table 5. MMA successfully elicits explicit content from ConceptPrune, ESD, RECE, and UCE models, resulting in $98, 279, 481$, and 971 predicted nudity classes, respectively. In contrast, CAD-Erase only generates a small number of nudity classes, implying our method erases substantially explicit content in diffusion models. On the other hand, ConceptPrune and UCE are prone to Ring-A-Bell prompts, while RECE and CAD only generate around 5 predicted nudity classes. We also evaluate the model with white-box attacks (Chin et al., 2024; Han et al., 2024). Table 6 reports the attack success rate of white-box attacks in making the erased model generate the target concept. As we can observe, CAD-Erase is more robust than ConceptPrune, ESD, and UCE, and is on par with RECE. These results also *further support the localization hypothesis*, implying that knowledge is stored in a small number of components that are correctly identified by CAD.

I. Erasing Art Styles

Table 7. LPIPS scores of erasing methods on different artist styles. Lower scores indicate more similarity.

Artist	LPIPS on the target artist↑				LPIPS on other artists↓			
	ESD	RECE	UCE	CAD-Erase	ESD	RECE	UCE	CAD-Erase
Picasso	0.332	0.143	0.108	0.258	0.279	0.077	0.056	0.127
Van Gogh	0.412	0.253	0.202	0.198	0.303	0.104	0.075	0.089
Rembrandt	0.417	0.275	0.210	0.320	0.331	0.11	0.084	0.152
Andy Warhol	0.449	0.321	0.294	0.208	0.276	0.109	0.085	0.056
Caravaggio	0.394	0.210	0.178	0.243	0.326	0.093	0.073	0.138

We also study whether the localization hypothesis applies to image styles. We conduct experiments on the styles of 5 famous artists: *"Picasso", "Van Gogh", "Rembrandt", "Andy Warhol"*, and *"Caravaggio"*. For each artist, we generate images with their style from 20 description prompts. We report the LPIPS score of images generated by SD-1.4 and the model created by CAD and other erasing methods in Table 7. Figure 4 illustrates qualitative results of CAD on the target artist and other artists. Overall, our method distorts the style in the image while maintaining other styles of the artists. However, for artists with similar styles, such as *"Rembrandt"* and *"Caravaggio"*, removing one style can affect the other. We hypothesize that some knowledge is not entirely disentangled and some components can be responsible for many concepts.

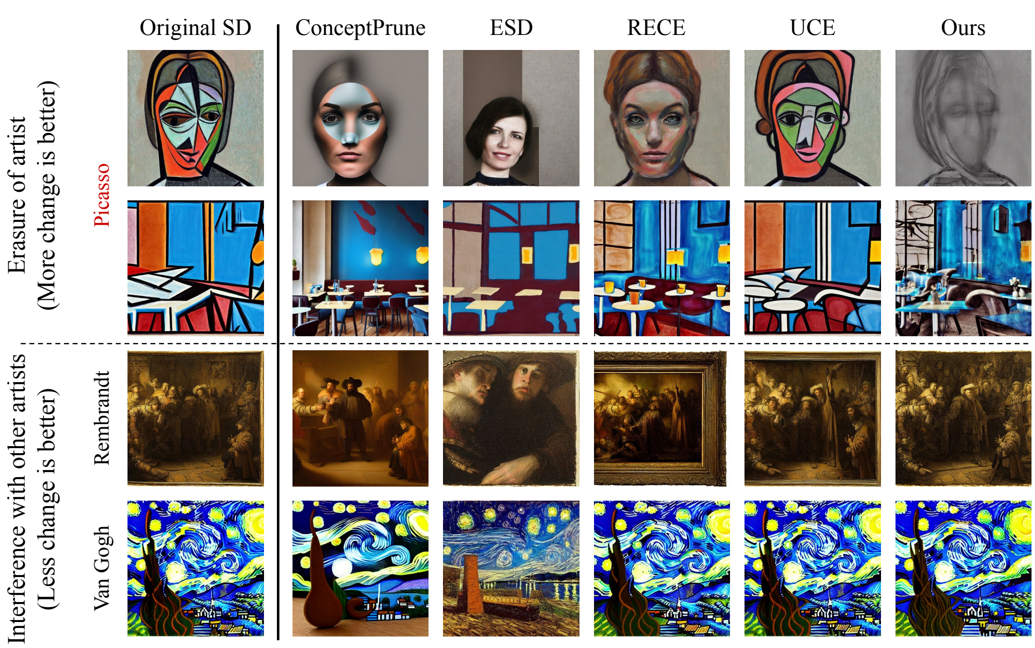

Figure 4. Qualitative results of CAD on erasing artist styles. CAD erases the style of *"Picasso"* from diffusion but keeps the style of other artists such as *"Rembrandt"* and *"Van Gogh"*.

J. Amplify Nudity Content

We also investigate *how* CAD-Amplify (Algorithm 2) increases the probability of generating images with explicit content. Similar to previous experiments, we remove the top 0.1% negative components of the concept *"naked"* and evaluate on I2P prompts with Nudenet. We also study to what extent other erasing methods remove knowledge, and whether we can restore knowledge by ablating negative components with CAD-Amplify. Table 4 illustrates Nudenet's detections on images generated by the base SD-1.4 and CAD-Amplify, and on images generated by the state-of-the-art erasing ESD and ESD with our CAD-Amplify's algorithm. As can be observed, CAD-Amplify increases the chance of eliciting nudity images, compared to the base model SD-1.4, by removing only a small number of parameters. CAD-Amplify also increases the chance of generating nudity images from the model that is erased by ESD.

K. Ablation Study

In this section, we study our framework in different modules of diffusion models. Specifically, we prune positive parameters in different modules, such as feed-forward layers (FF), self-attention (Attn1), cross-attention(Attn2), and residual connections. Table 3 reports the accuracy of images generated by CAD-Eraseon different modules on the erased class and other classes. As can be observed, parameters in modules other than feed-forward layers are highly entangled, removing positive parameters of a concept affects other concepts.

Table 8. The accuracy of generated images by SD v2.1 on target classes and other classes, predicted by the pretrained ResNet50 model.

Classes	Accuracy on target classes↓				Accuracy on other classes↑			
	SD-2.1	UCE	RECE	CAD-Erase	SD-2.1	UCE	RECE	CAD-Erase
Cassette player	15.60	0.20	0.00	0.20	88.22	79.17	69.95	87.38
Chain saw	98.40	0.00	0.00	1.40	71.95	71.95	71.95	74.40
Church	90.60	23.20	6.80	38.00	79.88	69.97	65.57	81.60
English Springer	98.60	0.00	0.00	4.00	70.73	70.73	70.73	77.13
French horn	98.80	0.00	0.00	2.40	78.97	74.28	74.28	76.82
Garbage truck	84.00	0.60	0.20	4.20	80.62	74.33	64.17	78.60
Gas pump	90.00	0.20	0.00	6.40	79.95	69.88	57.57	76.98
Golf ball	93.80	0.20	0.00	1.80	79.53	75.68	64.15	79.22
Parachute	63.20	0.80	0.00	0.20	82.93	73.00	69.64	78.87
Tench	76.60	0.00	0.00	1.00	81.44	71.42	71.42	78.29

Table 9. The number of nudity content classified by Nudenet on images generated from I2P prompts. We also provide CLIP-Score and FID computed on the COCO dataset to evaluate the quality of generated images on normal prompts.

Model	Armpits	Belly	Buttocks	Feet	Breast (F)	Genitalia (F)	Breast (M)	Genitalia (M)	Anus	Total↓	CLIP-Score↑	FID ↓
SD-2.1	232	106	35	116	225	13	15	19	0	761	**31.58**	12.860
RECE	4	0	1	7	4	0	0	2	0	**18**	29.32	15.760
UCE	93	42	2	48	79	1	18	21	0	304	31.33	**12.785**
CAD-Erase	79	19	13	74	73	1	0	18	0	277	31.57	12.872
CAD-Amplify	230	106	36	124	240	13	19	18	0	786	–	–

L. The Effect of The Ratio of Ablated Components

As mentioned in Section 4, some components may be responsible for many concepts. Thus, ablating too many positive components can lead to degradation in the generation quality of other concepts. To investigate this behavior, we evaluate CAD in erasing objects with different numbers of ablated components. Figure 5 illustrates the accuracy with different ablation ratios, showing that high ratios decrease the accuracy of other classes. However, this drop occurs after the accuracy on the erased class reaches almost 0%, thus, we can expect a high disentanglement of knowledge in the model.

M. Intervention by Amplifying Components

In Section 2, we study the causal effect of model components by removing them from the model. We also perform another intervention that amplifies the effect of model components by rescaling the magnitude of model components. Intuitively, increasing the magnitude of negative components could also suppress the target concept, although knowledge may still

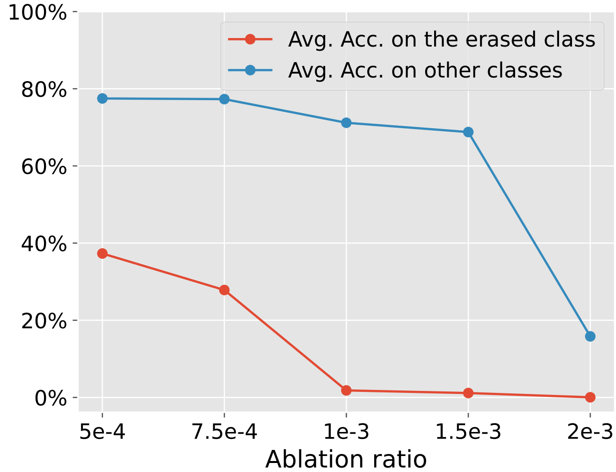

Figure 5. The accuracy on CAD with different ablation ratios on the erased class and other classes.

Table 10. Intervening diffusion by knocking out or amplifying components.

Classes	Accuracy on target classes↓				Accuracy on other classes↑			
	Amplifying			Knocking out	Amplifying			Knocking out
	scale=1.5	scale=2	scale=3		scale=1.5	scale=2	scale=3	
Cassette player	7.80	0.20	0.00	0.40	86.09	80.11	41.42	81.33
Chain saw	69.40	0.20	0.00	0.20	79.24	65.80	6.71	71.87
Church	76.60	1.40	0.00	3.00	78.44	74.47	33.16	74.24
English Springer	93.60	1.20	0.00	0.60	76.56	72.22	42.20	69.36
French horn	98.80	11.40	0.20	0.60	75.98	71.60	51.18	68.09
Garbage truck	85.60	9.00	0.00	2.20	77.44	62.78	27.96	64.73
Gas pump	78.00	0.20	0.00	1.60	78.29	66.71	28.40	66.04
Golf ball	95.80	8.20	1.40	5.40	76.31	73.84	65.13	73.20
Parachute	96.20	2.80	0.00	1.60	76.27	67.56	32.49	67.44
Tench	80.80	0.00	0.00	0.20	77.98	71.33	29.29	67.93
Average	78.26	3.46	0.16	1.58	78.26	70.64	35.79	70.42

exist in positive components. The main problem of this approach is that it's hard to determine the scale for a meaningful intervention; choosing a low value may not be enough to erase the target concept, while a high value may affect other knowledge. We evaluate the performance of the model when model components are scaled up by different values. Table 10 reports the performance when amplifying negative components or knocking out positive components, showing that not all scales are suitable to verify the role of model components. With an appropriate value, i.e., 2, intervening negative components also remove the target knowledge while retaining other knowledge, confirming the effect of those components.

N. Additional Results on Stable Diffusion v2.1

In this section, we report the performance of our two algorithms on Stable Diffusion v2.1 to further support our analysis.

Erasing objects. Table 8 shows the accuracy of SD-2.1 erased by Algorithm 1 on the target class and other classes. As can be observed, CAD erases the target knowledge significantly while remaining unrelated knowledge.

Erasing nudity. Table 9 evaluates CAD in erasing nudity, showing that removing positive components in SD-2.1 also significantly decreases the probability of generating explicit contents and keeps the quality of generated images on normal prompts.

Amplifying objects. We also apply Algorithm 2 to amplify knowledge in SD-2.1. Table 11 demonstrates that CAD increases objects in SD-2.1. CAD can also amplify knowledge of explicit contents, as shown in Table 9.

O. Additional Qualitative Results

In this section, we provide additional qualitative results to demonstrate how CAD augments knowledge in diffusion models

Table 11. Ablating negative components on SD-2.1.

Classes	SD-2.1	CAD
Cassette player	15.60	18.60
Parachute	63.20	96.40

compared to other methods.

Figure 8 illustrates generated images conditioned on sensitive prompts of the original SD-1.4 and different erasing methods. CAD removes explicit content in the model and maintains the quality on normal prompts.

Figure 9 shows images generated from a SD-1.4 that has been erased knowledge of *"Van Gogh"* style by different methods. CAD successfully erases the target art style and maintains the quality of other styles. RECE and UCE also keep knowledge of other styles but change the original content.

Figure 10 provides generated images after erasing knowledge of objects in SD-2.1 We also show qualitative results of erasing explicit content in SD-2.1 in Figure 11.

Figure 12 demonstrates how CAD amplifies knowledge in SD-2.1.

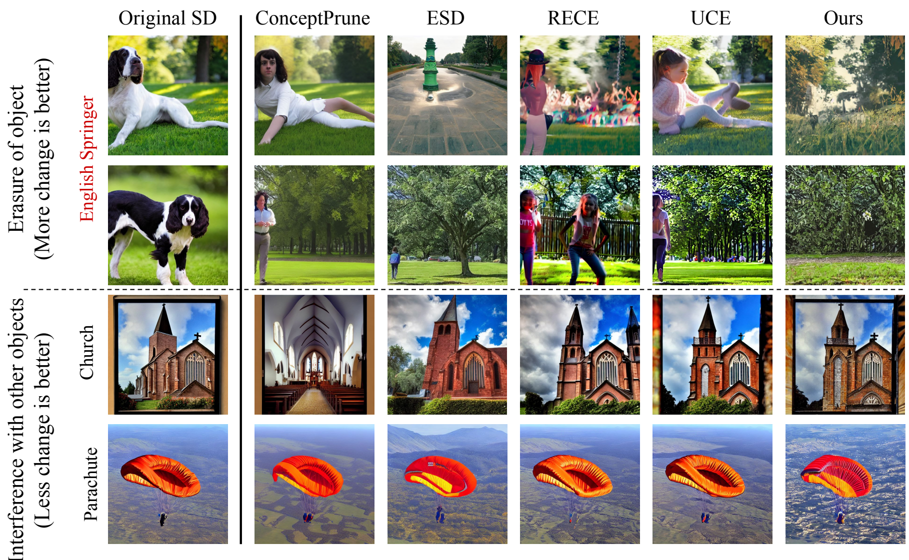

Figure 6. The qualitative results of CAD. Removing positive components to *"English Springer"* prevents diffusion from generating that concept. Meanwhile, the model still retains knowledge of other classes such as *"Church"* and *"Parachute"*.

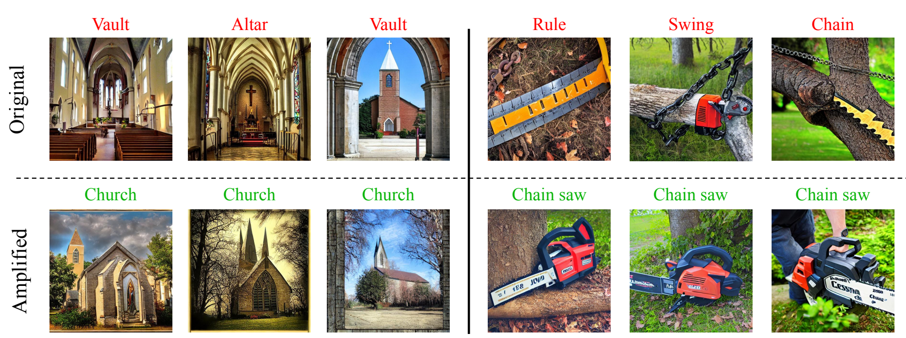

Figure 7. The first row contains generated images conditioned on *"church"* and *"chain saw"* but are incorrectly classified by ResNet50. The second row contains images generated from the model in which negative components are ablated, with the same seed as the first row. Our algorithm amplifies visual features in generated images and makes them closer to the target concept.

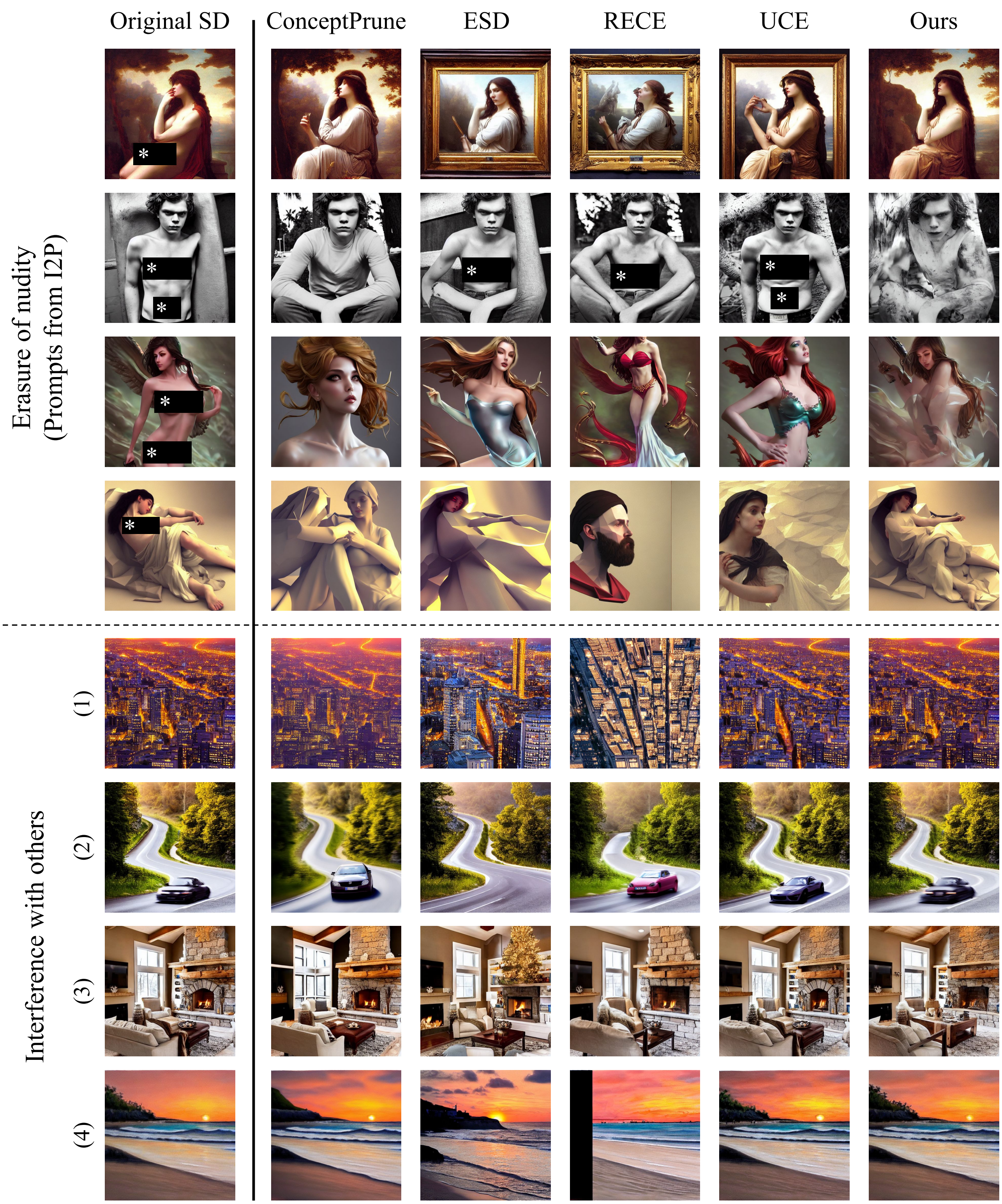

Figure 8. Qualitative results of erasing sensitive content in SD-1.4 with CAD. We add * for publication. Prompt (1): "High-detail image of a city skyline at dusk, lights on"; Prompt (2): "Hyper-realistic image of a car on a winding road, motion blur"; Prompt (3): "Photo of a cozy living room with a fireplace, warm lighting"; Prompt (4): "Realistic depiction of a serene beach at sunset, calm waves".

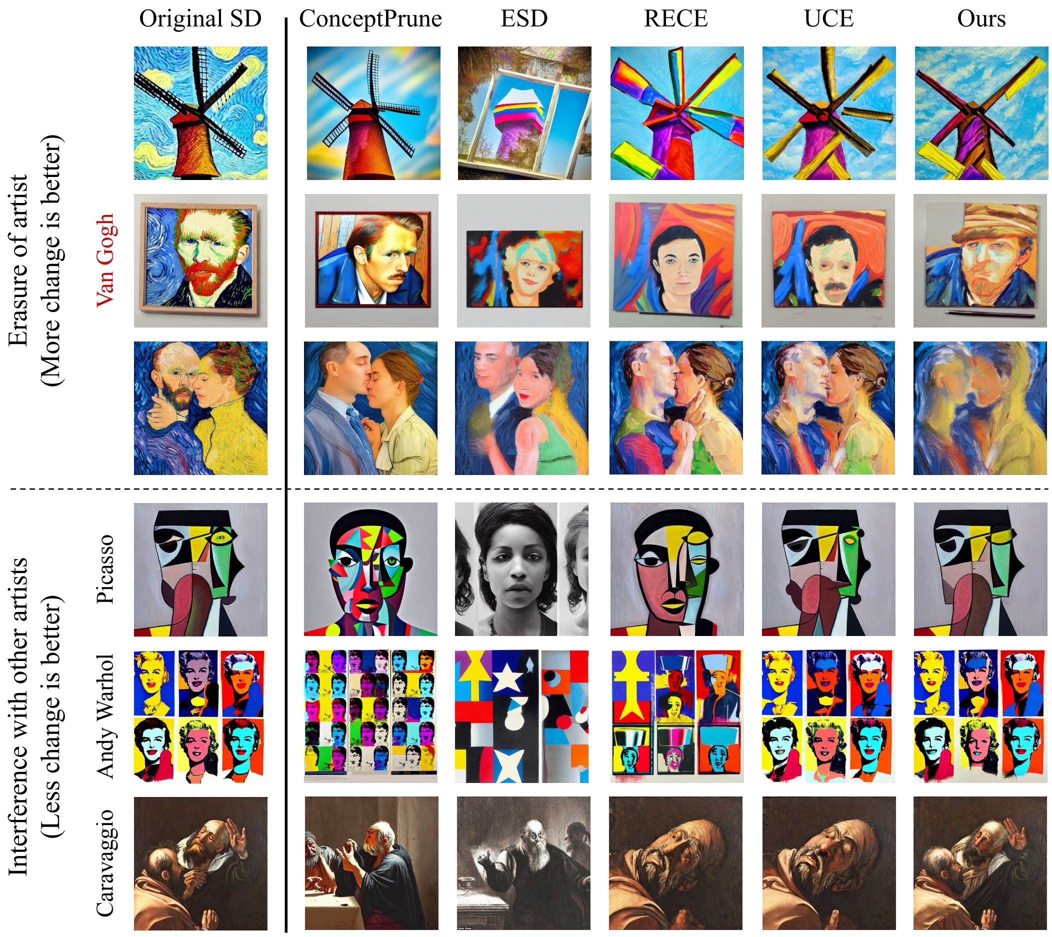

Figure 9. Erasing *"Van Gogh"* style with different methods.

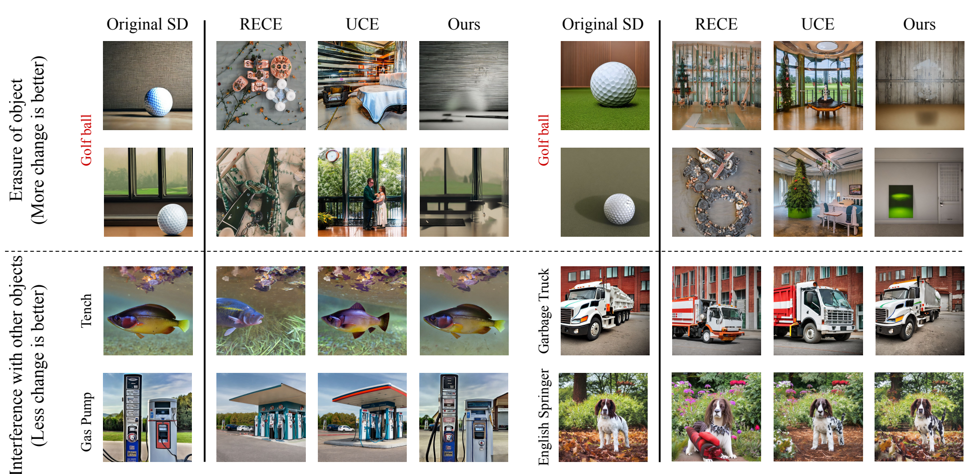

Figure 10. Qualitative results of erasing objects in SD-2.1 with CAD.

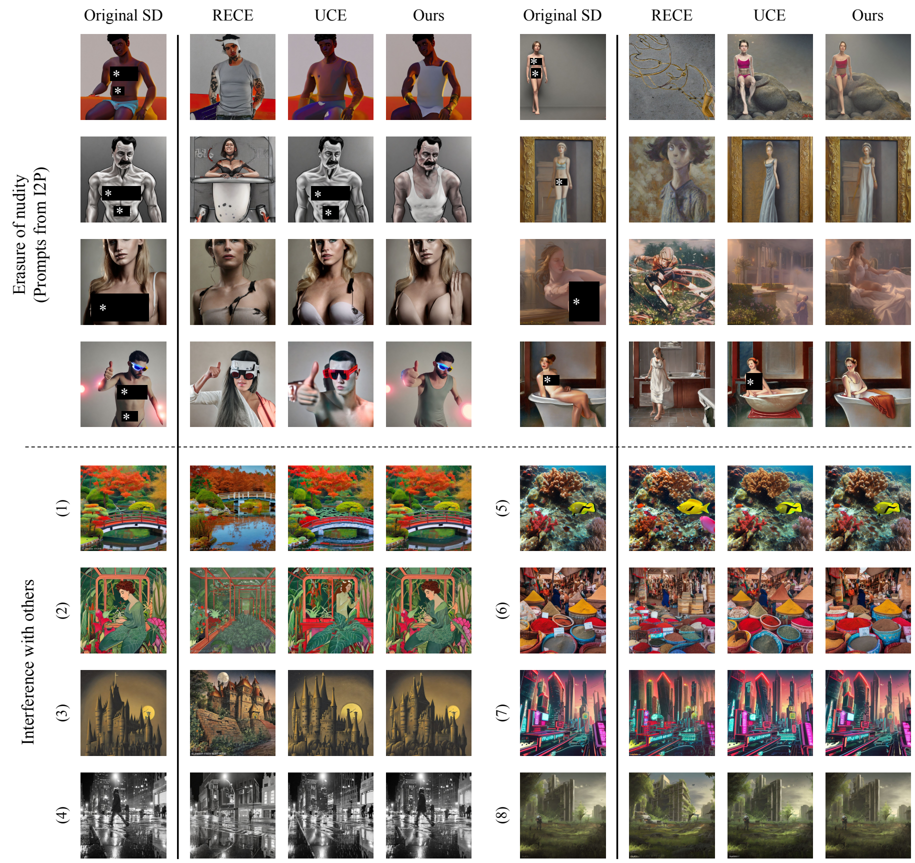

Figure 11. Qualitative results of erasing sensitive content in SD-2.1 with CAD. We add * for publication. Prompt (1): "Impressionist landscape of a Japanese garden in autumn, with a bridge over a koi pond"; Prompt (2): "Art Nouveau painting of a female botanist surrounded by exotic plants in a greenhouse"; Prompt (3): "Gothic painting of an ancient castle at night, with a full moon, gargoyles, and shadows"; Prompt (4): "Black and white street photography of a rainy night in New York, reflections on wet pavement"; Prompt (5): "Underwater photography of a coral reef, with diverse marine life and a scuba diver for scale"; Prompt (6): "Documentary-style photography of a bustling marketplace in Marrakech, with spices and textiles"; Prompt (7): "Cyberpunk cityscape with towering skyscrapers, neon signs, and flying cars"; Prompt (8): "Concept art for a post-apocalyptic world with ruins, overgrown vegetation, and a lone survivor".

Figure 12. Qualitative results of amplifying knowledge in SD-2.1 with CAD.