

528 This Appendix provides additional details, analysis, and quantitative and qualitative results to support
529 the main paper. Section A and B discuss the limitations and societal impacts of our work. We report
530 experimental setups and hyperparameters in Section C. Section D shows the performance of CAD on
531 different modules. Section E discusses the results of CAD with different ablation ratios. Section F
532 studies the other type of intervention on model components. We present experimental results for
533 Stable Diffusion v2.1 in Section G and additional qualitative results in Section H.

534 A Limitations

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

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

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

552 B Societal Impacts

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

558 C Experimental Setup

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

- 563 • For Stable Diffusion v1.4:
 - 564 – ESD: We follow the setting in the original paper and fine-tune the UNet with a learning
565 rate of $1e - 5$. To compute the objective, we generate images of the target class with a
566 guidance scale of 3. The scale of negative guidance in the objective is set to 1.
 - 567 – UCE. We apply UCE across ten objects within the Imagenette class and for the artistic
568 styles of Picasso, Van Gogh, Rembrandt, Andy Warhol, and Caravaggio, including the
569 nudity concept. The method includes a “preserve” parameter in artist styles, which
570 retains styles not targeted for erasure. We follow that setting, by erasing only one artist
571 style at each checkpoint while keeping the rest.
 - 572 – RECE. This method continues to fine-tune models using checkpoints previously erased
573 by UCE. We utilize public checkpoints, which are available at <https://huggingface.co/ChaoGong/RECE>. These checkpoints include models fine-tuned to erase concepts
574 such as nudity and Van Gogh style, besides 5 objects such as church, garbage truck,
575 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.

Table 7: 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

D 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 7 reports the accuracy of images generated by CAD-Erase on 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.

E The Effect of The Ratio of Ablated Components

As mentioned in Section 6, 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 7 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.

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	–	–

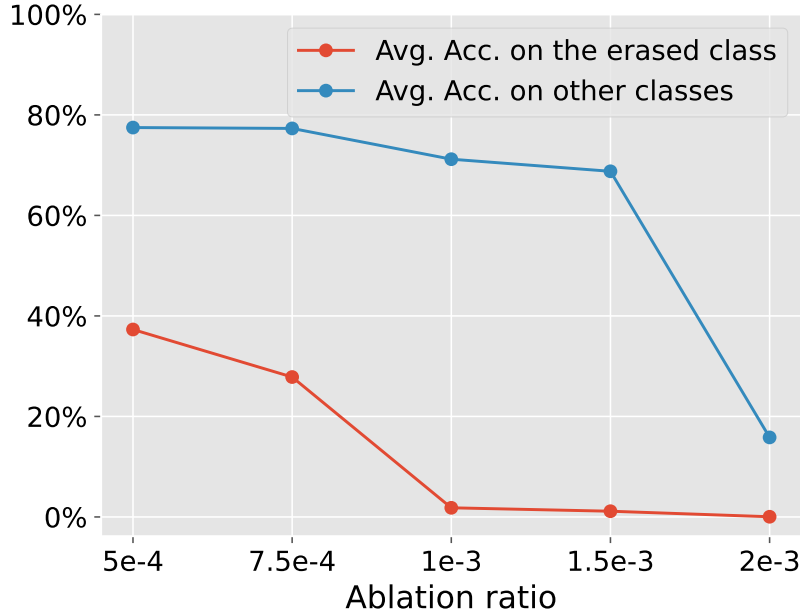


Figure 7: The accuracy on CAD with different ablation ratios on the erased class and other classes.

F Intervention by Amplifying Components

In Section 4, 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 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

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

624 components. With an appropriate value, i.e., 2, intervening negative components also remove the
625 target knowledge while retaining other knowledge, confirming the effect of those components.

626 G Additional Results on Stable Diffusion v2.1

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

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

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

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

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

638 H Additional Qualitative Results

639

640 In this section, we provide additional qualitative results to demonstrate how CAD augments knowl-
641 edge in diffusion models compared to other methods.

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

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

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

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

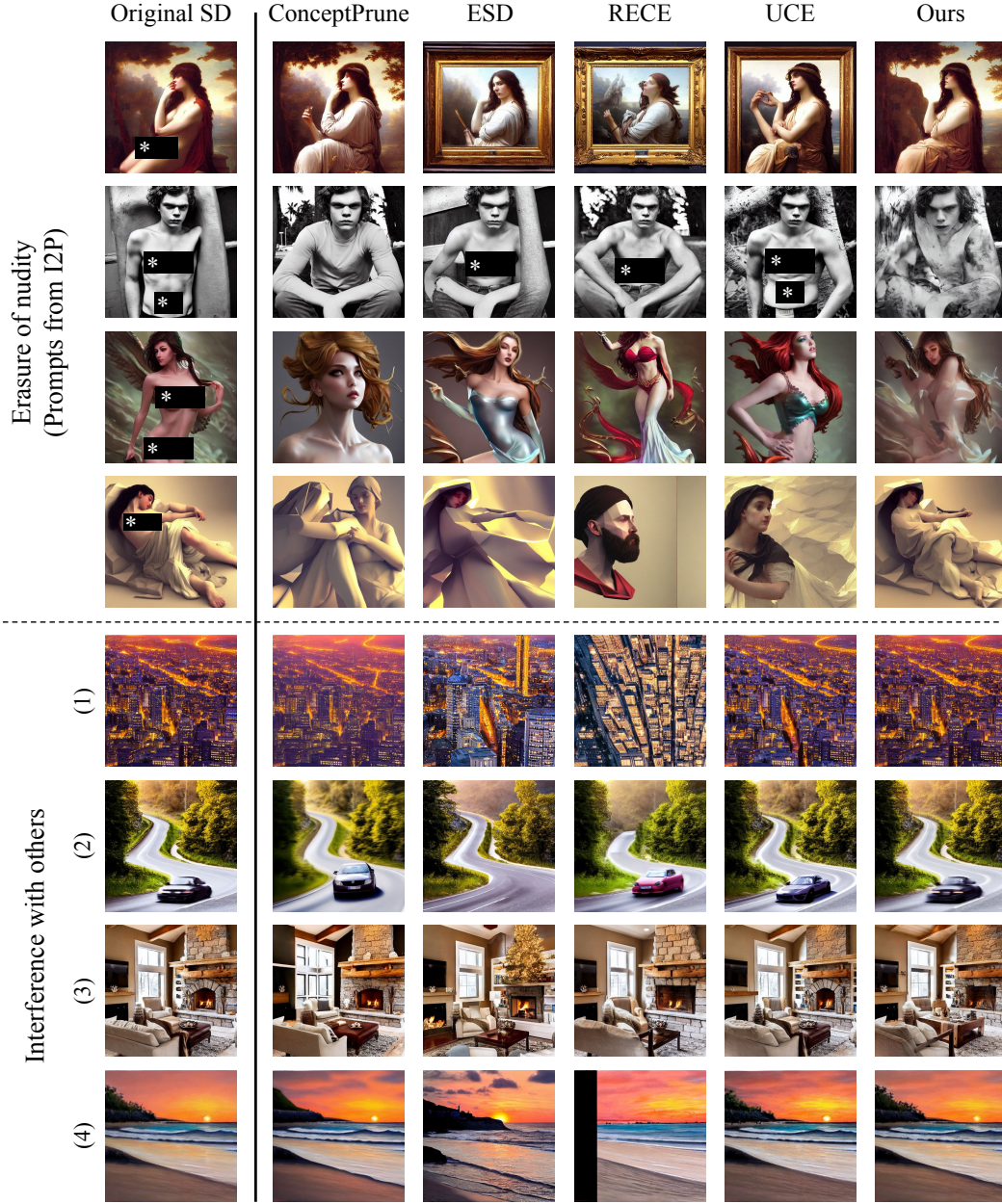


Figure 8: Qualitative results of erasing sensitive content in SD-1.4 with CAD. We add for publication. Prompt (1): “High-detail image of a city skyline at dusk, lights on”; Prompt (2): “Hyper-realistic image of a car on a winding road, motion blur”; Prompt (3): “Photo of a cozy living room with a fireplace,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



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

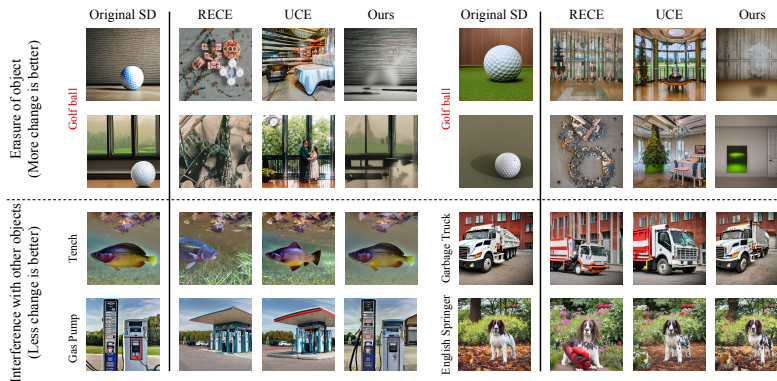


Figure 10: Qualitative results of erasing objects in SD-2.1 with CAD.

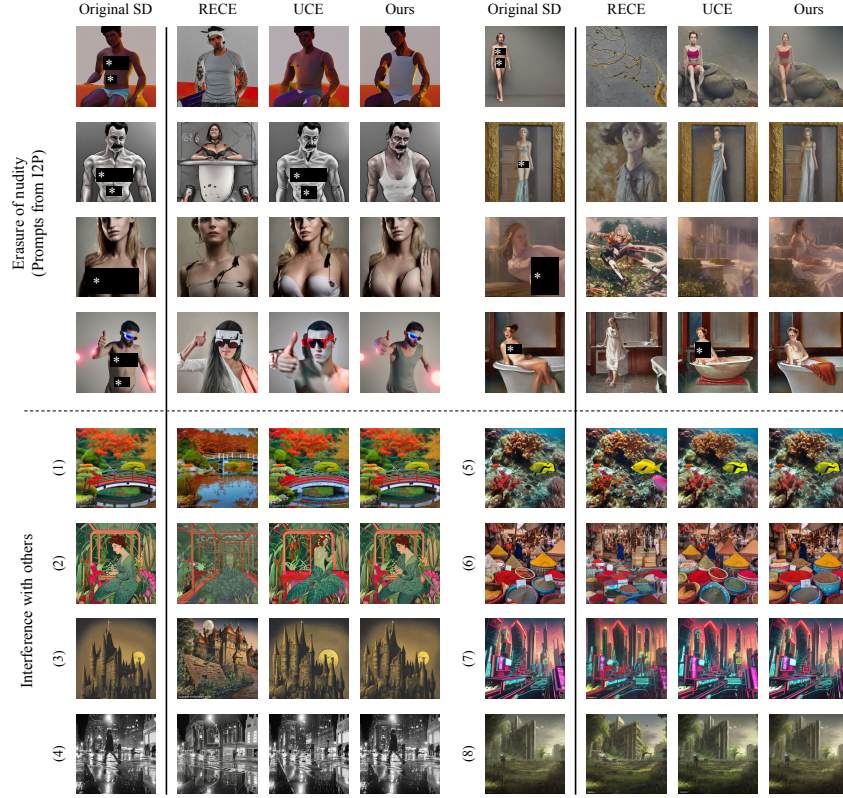


Figure 11: Qualitative results of erasing sensitive content in SD-2.1 with CAD. We add * for publication. Prompt (1): “Impressionist landscape of a Japanese garden in autumn, with a bridge over a koi pond”; Prompt (2): “Art Nouveau painting of a female botanist surrounded by exotic plants in a greenhouse”; 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.