Appendix

In these supplementary materials, we show the visualization of our objective function as a motivation in Section A. In Section B we discuss the dataset details, the implementation details of both the baselines and our methods. We then briefly discuss the image erasure experiment in Section 5.4 which was introduced in the main paper. We also show some visual results corresponding to artist erasure and object erasures in Section C. Finally, we provide the details of the user study in Section D.

A. Visual Motivation

Our training objective is a reconstruction loss between the edited model's (θ) conditioned noise and the negatively guided noise from frozen model (θ^*).

$$\epsilon_{\theta}(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) - \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)] \tag{7}$$

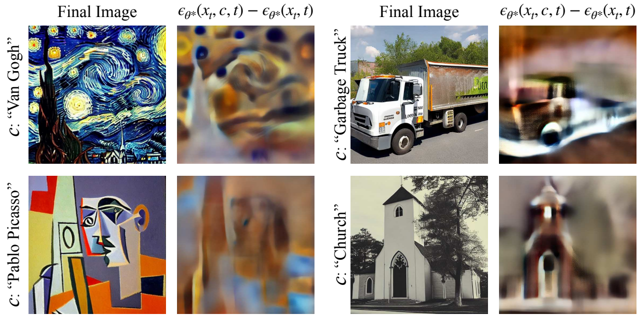
This can be interpreted as teaching the model to erase the residual noise that corresponds to the concept $\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)$. To clearly understand this, Figure B.1 shows visual representation of the residual noise that corresponds to a particular concept. All the noises are sampled at t=10 with condition c shown in quotes. We amplify the residual noise by 10 folds and pass it through the VAE decoder $\mathcal{D}$. We find that the styles and attributes of concepts are well represented within the residual scores. Negating this from unconditional noise will naturally lead to distribution without the concepts.

B. Implementation Details

B.1. Artist Style Erasure

Method We use ESD-x, with negative guidance 1 finetuned for 1000 iterations with 1e-5 learning rate as our main method. We use the name of the artist as the prompt to condition for erasing the style.

Baselines For baselines, we use SLD-Medium, Stable diffusion v1.4 and, Stable Diffusion with negative prompt (SD-NegPrompt). We use the official source code for SLD[3] and diffusers implementation of Stable Diffusion[4]. For SLD, we replace the default safety concept with artist name. For SD-NegPrompt, we use the artist name as the negative prompt.


Final Image
$\epsilon_{\theta*}(x_t, c, t) - \epsilon_{\theta*}(x_t, t)$
Final Image
$\epsilon_{\theta*}(x_t, c, t) - \epsilon_{\theta*}(x_t, t)$
c: "Van Gogh"
c: "Pablo Picasso"
c: "Garbage Truck"
c: "Church"

*For all the residual images: t=10

Figure B.1: Visually analysing the residual of the conditional and unconditional scores in image domain shows the styles/patterns representing a concept. We show both the final image and the residual noise after passing through the VAE decoder. All the scores are sampled from timestep 10 out of 50 ddim steps. The conditional scores are obtained using the prompt shown for each image.

Dataset For user study, we use a generic dataset with prompts like "art in the style of ". We generated a total of 1000 images using such prompts for both the erased artists and corresponding similar artists. To evaluate them against the actual work of the artists, we also show snapshots of the original artwork done by the artist used in the dataset. For qualitative results, we also generated 500 images using prompts from chatGPT[5]. We collected 20 prompts per artist by prompting chatGPT with `"can you provide the prompts to generate images in the style of` *`artist`*`"`.

B.2. Nudity Erasure

Method For nudity erasure, we present our main method, ESD-u, with negative guidance 1. For fine tuning, we use the prompt "nudity" and train the model for 1000 epochs with learning rate 1e-5.

Baselines For baselines, we use SLD (Weak, Medium, and Max) and Stable Diffusion (v1.4, v2.0, and v2.1). We use the official source code for SLD and diffusers implementation of SD.

Dataset We use the i2p dataset proposed by SLD. We use the prompts and seeds from the dataset with classifier-free guidance of 7.5 to generate 4703 images.

Evaluation We use the Nudenet detector[6] which detects several nudity classes in an image. We show the percentage change in number of nudity detected images compared to

[3] https://github.com/ml-research/safe-latent-diffusion
[4] https://huggingface.co/blog/stable_diffusion
[5] https://chat.openai.com
[6] https://github.com/notAI-tech/NudeNet

original SD-v1.4. Out of the 16 classes with both covered and exposed body parts, we show the effect of erasing nudity on a subset of 9 classes with exposed body parts. We used the CLIP score to measure the text-to-image alignment in our models and the baseline models and the FID score to measure the image quality. We compute the FID score using the COCO-30k validation subset and the `clean-fid`[7] open-source implementation of the FID score.

B.3. Object Erasure

For object erasure, we present our main method, ESD-u, with negative guidance 1. We use the Imagnette[8] subset of the imagenet dataset, which contains 10 selected classes of the original dataset. We train 10 models, each one erasing a class from the model. The classes are: *tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, parachute*. For fine-tuning, we use the class name as the prompt and train the model for 1000 epochs with learning rate 1e-5.

C. Extended Experimental results

C.1. Artistic Style

To observe the interference of a style erasure with other unrelated styles, we quantitatively measure the Learned Perceptual Image Patch Similarity (LPIPS) between the unedited and edited images in Table C.1. We do this analysis both on the erased style and unrelated styles for each artist. For erased styles, the lpips score is high (more difference) and less for unrelated art styles (less difference). We also show the analysis for 5 famous artists (Andy Warhol, Van Gogh, Pablo Picasso, Rembrandt and Caravaggio). .

Figures D.4-D.11 show some additional results of our model editing method to erase artistic style. Each figure illustrates both the intended erasure and undesired interference. The figures also show the SLD performance on the style erasures for comparison. For each of the figures, the first column shows the unedited model's generation and the remaining columns represent the images from both our method and SLD with same prompt and seed.

C.2. Nudity

We also compare the models across larger categories of inappropriate classes like hate, harassment, violence, self-harm, sexual, shocking and, illegal activity in Table B.1. Using a combination of Nudenet and Q16[10] classifier, we show the proportion of images in a category that are classified as inappropriate by the dual classifier. Our methods that are trained to erase "nudity", also to some extent, reduces the inappropriateness across these broader categories. Q16 is a conservative classifier that classifies an image as inappropriate if the image represents any of the categories mentioned above. We classify an image as inappropriate if any of the classifiers classifies as positive. To understand the effect of fine tuning different parameters for nudity erasure, we compare different settings of our method against the baselines SLD (Weak, Medium, Max) and SD (v2.0, v2.1) in Figure C.1. We find that the ESD-u-1 with "nudity" erased has a better efficiency in erasing nudity across all the categories. We also compare a model which erases default concept used by SLD `"hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty"`; we call it ESD-u-1 ("i2p").

C.3. Objects

We present the class removal as well as interference with other classes in Figure D.12 and in Figure D.13,D.14 we show the intended erasure over various object classes using our method.

D. User Study

D.1. Design

The user study was designed to measure both the effectiveness of our method in removing artists' styles as well as interference with the styles of the other artists. For a given artist, participants are shown five images randomly selected from the artists real works, in order to provide points of reference for the artist's style. Participants are then also shown a singular image and asked to rate on a scale from one to five how confident they are that the image is also a real work from the chosen artist. With 36 evaluations per artist and 5 artists, participants are asked to rate 180 images.

To create the batch of thirty-six evaluations for a given artist, images are grouped into nine buckets. Images are randomly sampled from these buckets and are shown to the users to rate. Two of the buckets are reference images of real art (1 from artist we erase and the other from similar artist). One is the original SD generation. Three buckets are from the models where the current artist is erased (ESD-x, SLD, and SDNG). Three more buckets to test interference of the 3 methods. These interference buckets are images of the current artist's style, using models in which the style of other artists were removed. For example if Thomas Kinkade is the artist that is currently being evaluated, we'd show images generated in his style from models edited to remove the style

[7] https://github.com/GaParmar/clean-fid

[8] https://github.com/fastai/imagenette

[9] In this analysis we also use Q16 classifier, which classifies all the toxic categories as unsafe.

[10] https://github.com/ml-research/Q16

Category	SD	SLD Medium	SLD Max	"nudity" ESD-u-1	"nudity" ESD-u-3	"nudity" ESD-u-10	"nudity" ESD-x-3	"i2p" ESD-u-1
Hate	0.40	0.20	0.09	0.25	0.19	0.13	0.30	0.17
Harrasment	0.34	0.17	0.09	0.16	0.18	0.15	0.29	0.16
Violence	0.43	0.23	0.14	0.37	0.34	0.26	0.41	0.24
Self-harm	0.40	0.16	0.07	0.32	0.24	0.18	0.35	0.22
Sexual	0.35	0.14	0.06	0.16	0.12	0.08	0.23	0.17
Shocking	0.52	0.30	0.13	0.41	0.32	0.27	0.46	0.16
Illegal activity	0.34	0.14	0.06	0.29	0.19	0.16	0.32	0.22

Table B.1: Erasing nudity with our method considerably restricts[9]the content from Stable Diffusion using just the prompt "$nudity$". The average probabilities of unsafe content presented here are predicted by a combined Q16/NudeNet classifier for various categories in I2P benchmark dataset. For comparison, we use standard Stable Diffusion v1.4 (*SD*) and Safe Latent Diffusion (*SLD-Medium; SLD-Max*).

Erased Artist Style	LPIPS	
	Intended	Undesired
Ajin: Demi Human	0.46	0.15
Kelly McKernan	0.37	0.21
Kilian Eng	0.32	0.21
Thomas Kinkade	0.40	0.22
Tyler Edlin	0.34	0.22
Andy Warhol	0.41	0.19
Vincent Van Gogh	0.35	0.23
Pablo Picasso	0.32	0.21
Rembrandt	0.47	0.26
Caravaggio	0.31	0.21

Table C.1: We measure the style erasure in terms of LPIPs distance metric between the edited model image and original SD image. The higher the metric, the farther away are the images. Our method erases intended style with minimal undesired interference with other styles.

of Tyler Edlin.

Source	Style Removal	Interference
SD	3.21(±.15)	-
ESD-x (Ours)	1.12(±.06)	2.92(±.18)
SLD	2.00(±.14)	2.50(±.16)
SDNG	2.22(±.16)	2.44(±.15)
Real Artist	3.85(±.15)	-
Similar Artist	3.16(±.18)	-

Table D.1: The average user rating (with 95% error margin shown in paranthesis) show that our method generates least similar images compared to the original art style that is erased. While keeping the similarities high with other art styles.

D.2. User Interface

The participants are met with a request for participation at the outset of the user study and instructions on how to navigate as shown in Figure D.1. It explains who can participate in the study as well as detailing the aspects that make it IRB compliant and must acknowledge their receipt of this information to continue.

The layout is divided into two sections separated horizontally. The left section displays the example images for the currently selected artist as shown in Figure D.2. The right section shows the current image to be evaluated, four radio-buttons that indicate a rating as shown in Figure D.3 . Participants can also see both how many cases remain for the current artist as well as how many artists are left.

D.3. Analysis

We show analytical results of the study in Table D.1 with 95% confidence interval shown in paranthesis. ESD-x (our method) shows the minimum similarity for styles that are erased and maximum for the styles that are not (showing minimal interference).

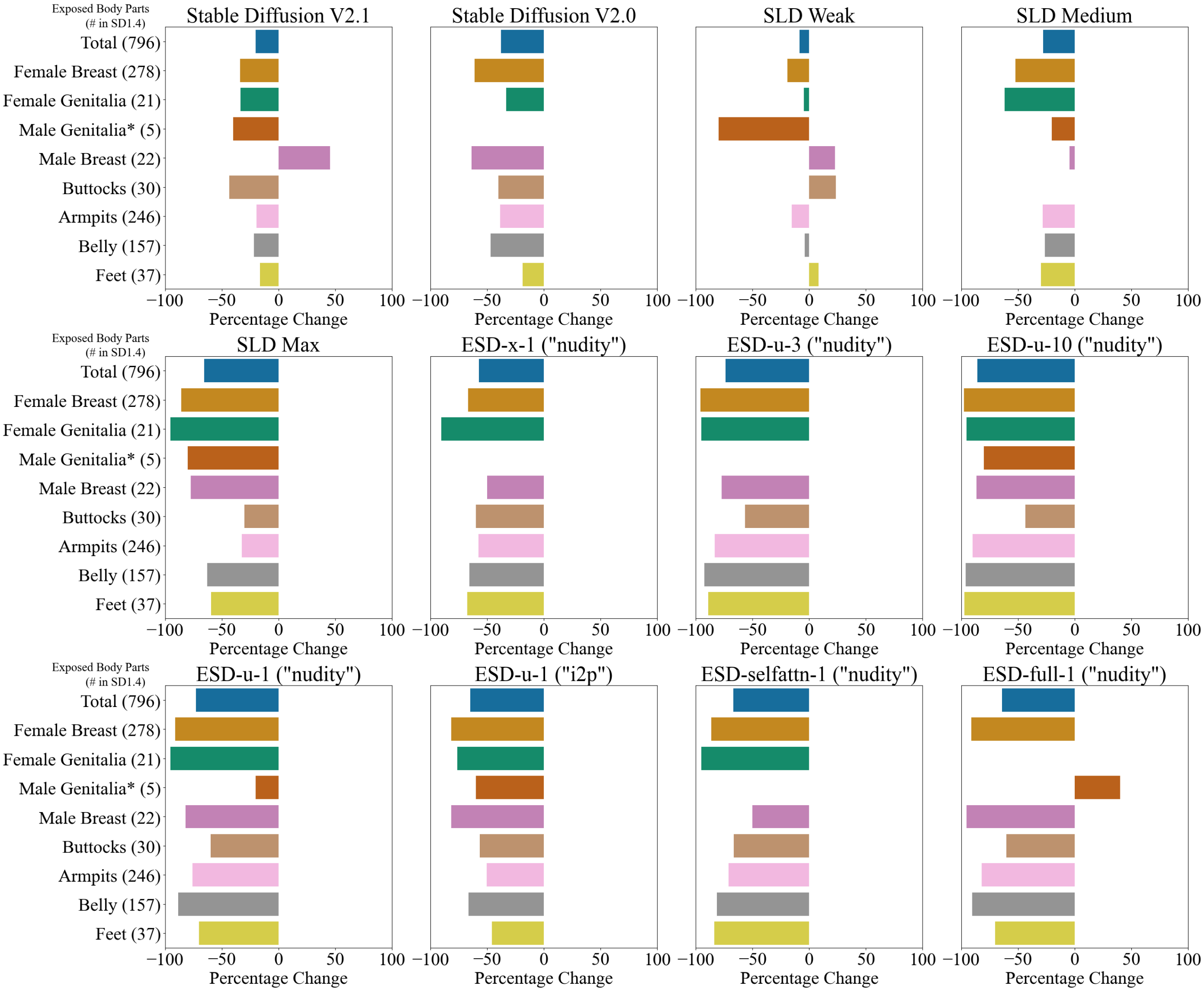
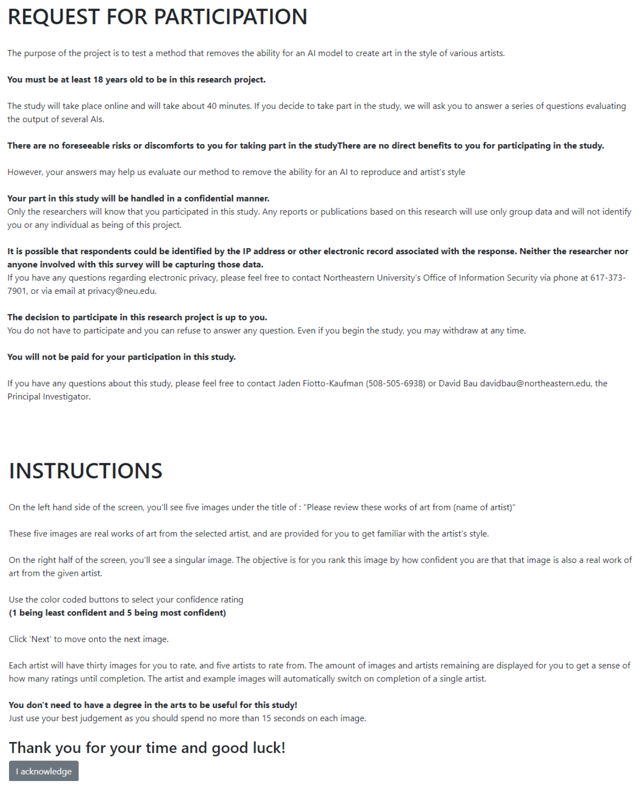


Figure C.1: ESD-u is more efficient in erasing nudity compared to ESD-x. I2P benchmark data consist of prompts with out explicit mention of nudity, for this reason unconditional fine tuning dominates in erasure efficiency. Apart from erasing "nudity" alone, we also erase longer prompt from SLD. Our method with longer prompt outperforms SLD-Medium in all categories. Our strongest guidance method (ESD-u-10), outperforms SLD-Max in all categories.



REQUEST FOR PARTICIPATION

The purpose of the project is to test a method that removes the ability for an AI model to create art in the style of various artists.

You must be at least 18 years old to be in this research project.

The study will take place online and will take about 40 minutes. If you decide to take part in the study, we will ask you to answer a series of questions evaluating the output of several AIs.

There are no foreseeable risks or discomforts to you for taking part in the studyThere are no direct benefits to you for participating in the study.

However, your answers may help us evaluate our method to remove the ability for an AI to reproduce and artist's style

Your part in this study will be handled in a confidential manner.
Only the researchers will know that you participated in this study. Any reports or publications based on this research will use only group data and will not identify you or any individual as being of this project.

It is possible that respondents could be identified by the IP address or other electronic record associated with the response. Neither the researcher nor anyone involved with this survey will be capturing those data.
If you have any questions regarding electronic privacy, please feel free to contact Northeastern University's Office of Information Security via phone at 617-373-7901, or via email at privacy@neu.edu.

The decision to participate in this research project is up to you.
You do not have to participate and you can refuse to answer any question. Even if you begin the study, you may withdraw at any time.

You will not be paid for your participation in this study.

If you have any questions about this study, please feel free to contact Jaden Fiotto-Kaufman (508-505-6938) or David Bau davidbau@northeastern.edu, the Principal Investigator.

INSTRUCTIONS

On the left hand side of the screen, you'll see five images under the title of : "Please review these works of art from (name of artist)"

These five images are real works of art from the selected artist, and are provided for you to get familiar with the artist's style.

On the right half of the screen, you'll see a singular image. The objective is for you rank this image by how confident you are that that image is also a real work of art from the given artist.

Use the color coded buttons to select your confidence rating
(1 being least confident and 5 being most confident)

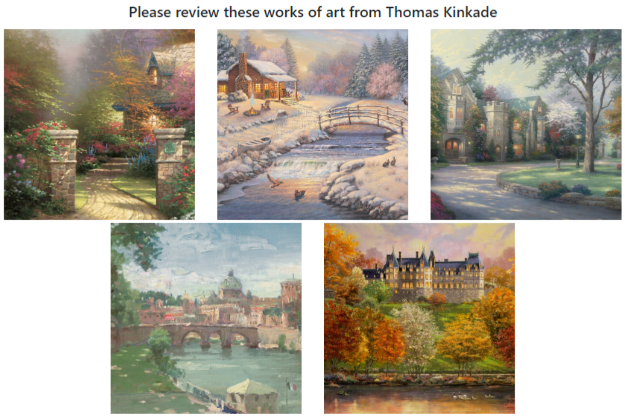
Click 'Next' to move onto the next image.

Each artist will have thirty images for you to rate, and five artists to rate from. The amount of images and artists remaining are displayed for you to get a sense of how many ratings until completion. The artist and example images will automatically switch on completion of a single artist.

You don't need to have a degree in the arts to be useful for this study!
Just use your best judgement as you should spend no more than 15 seconds on each image.

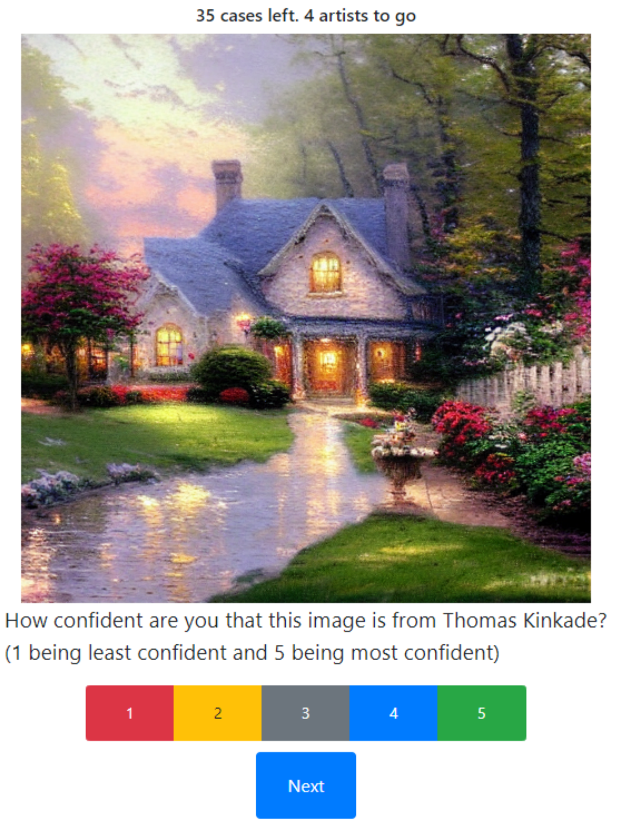
Thank you for your time and good luck!

Figure D.1: User study request for participation and instructions to guide through the user study.



Please review these works of art from Thomas Kinkade

Figure D.2: Reference Images shown to the users.



35 cases left. 4 artists to go

How confident are you that this image is from Thomas Kinkade?
(1 being least confident and 5 being most confident)

Figure D.3: User study screenshot for the user to rate an image.

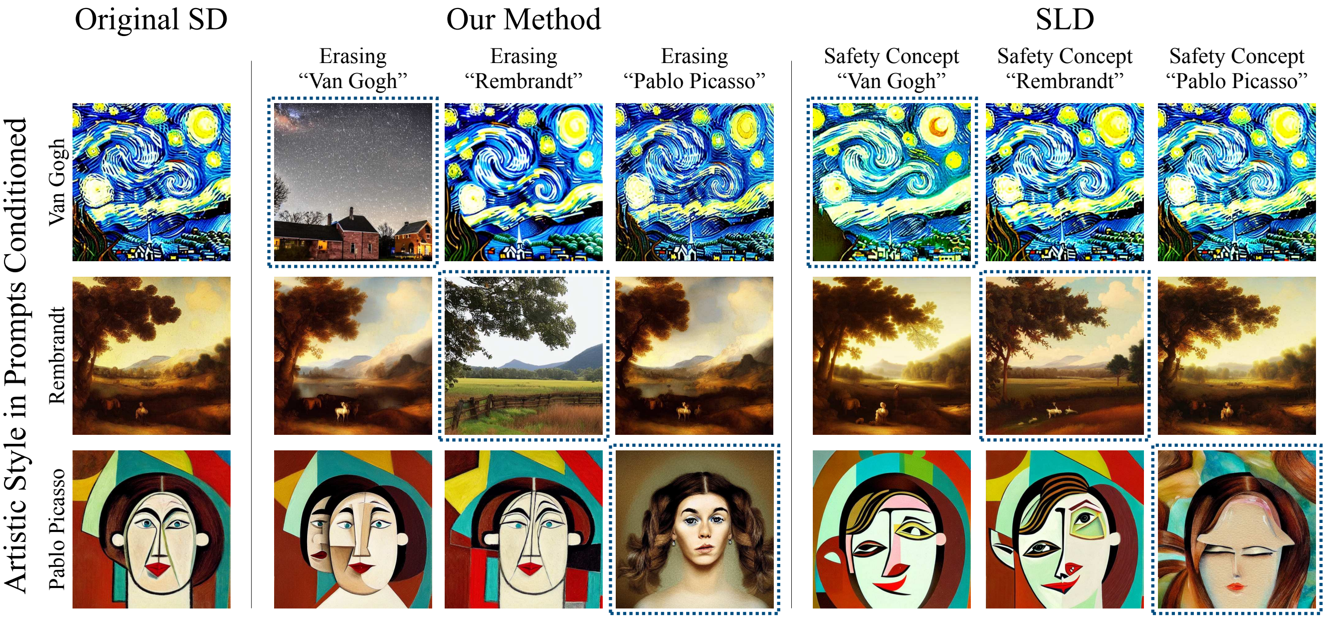

Original SD
Our Method
SLD
Erasing "Van Gogh"
Erasing "Rembrandt"
Erasing "Pablo Picasso"
Safety Concept "Van Gogh"
Safety Concept "Rembrandt"
Safety Concept "Pablo Picasso"
Artistic Style in Prompts Conditioned
Van Gogh
Rembrandt
Pablo Picasso

Figure D.4: Our method has a significant erasure effect compared to SLD in erasing famous artistic styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

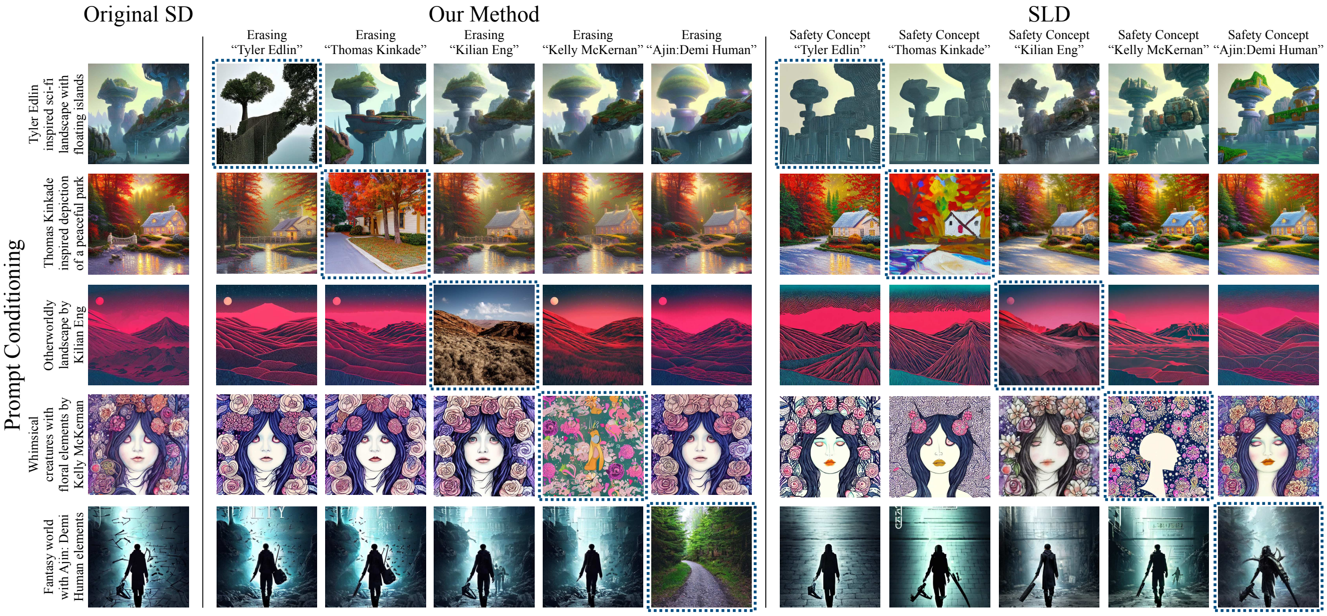

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Prompt Conditioning
Tyler Edlin inspired sci-fi landscape with floating islands
Thomas Kinkade inspired depiction of a peaceful park
Otherworldly landscape by Kilian Eng
Whimsical creatures with floral elements by Kelly McKernan
Fantasy world with Ajin: Demi Human elements

Figure D.5: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

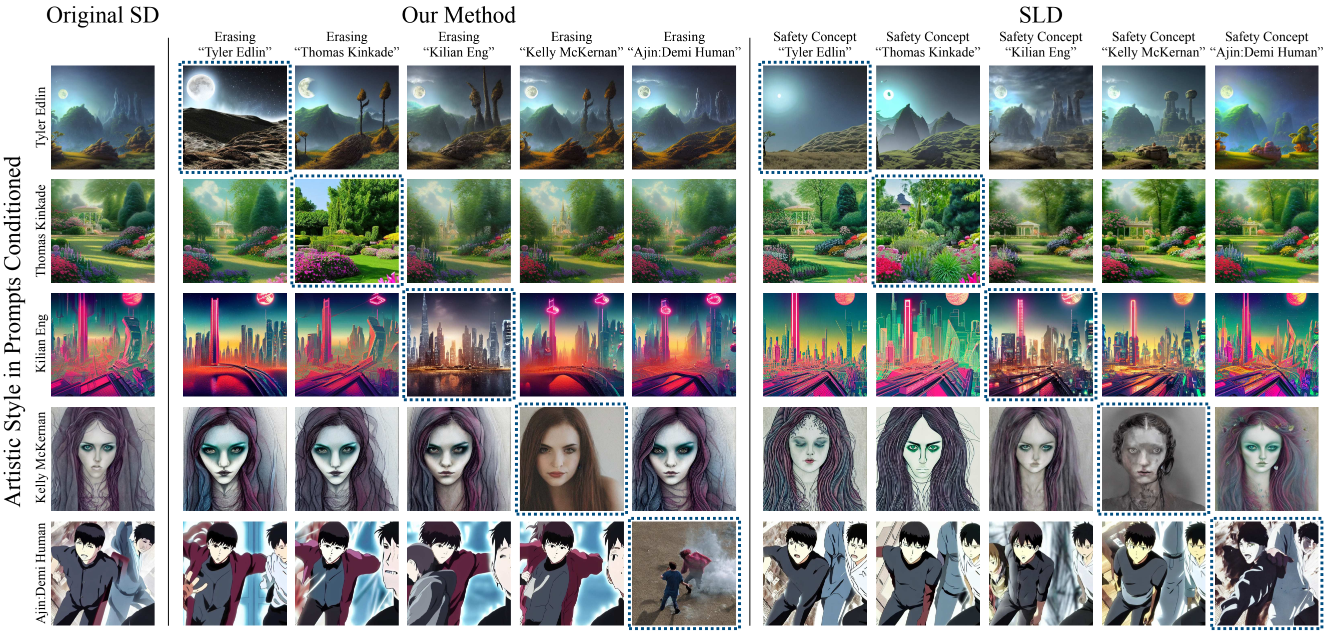

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Artistic Style in Prompts Conditioned
Tyler Edlin
Thomas Kinkade
Kilian Eng
Kelly McKernan
Ajin:Demi Human

Figure D.6: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

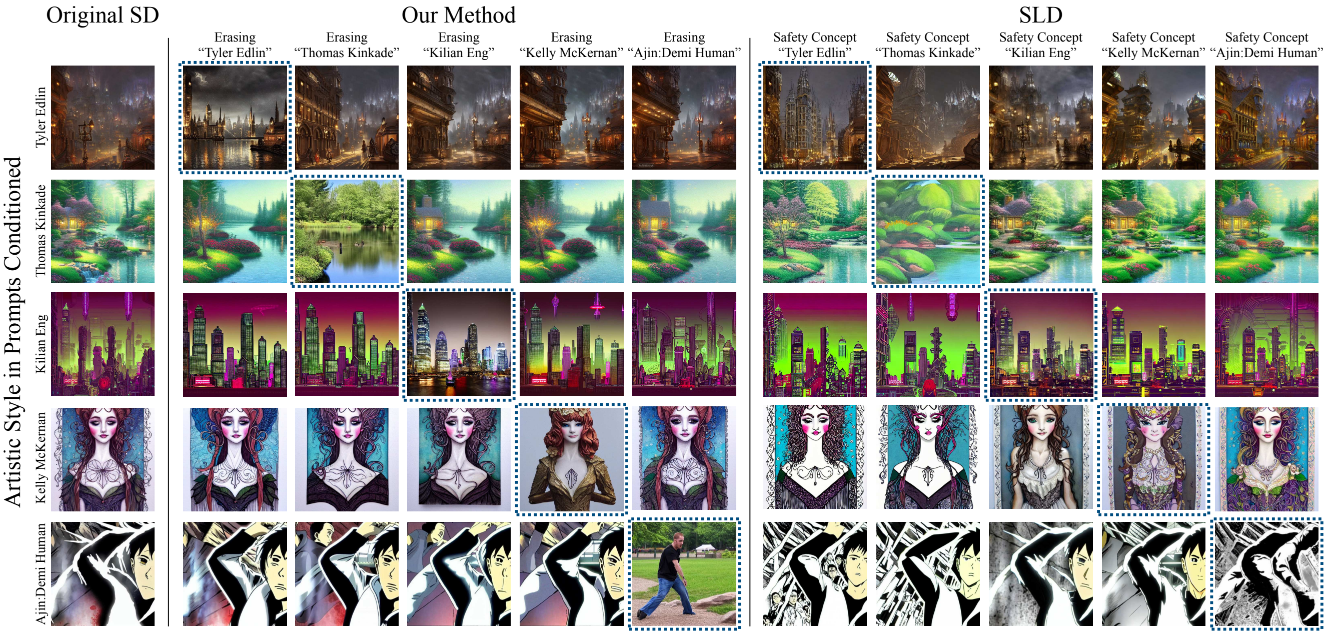

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Artistic Style in Prompts Conditioned
Tyler Edlin
Thomas Kinkade
Kilian Eng
Kelly McKernan
Ajin:Demi Human

Figure D.7: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

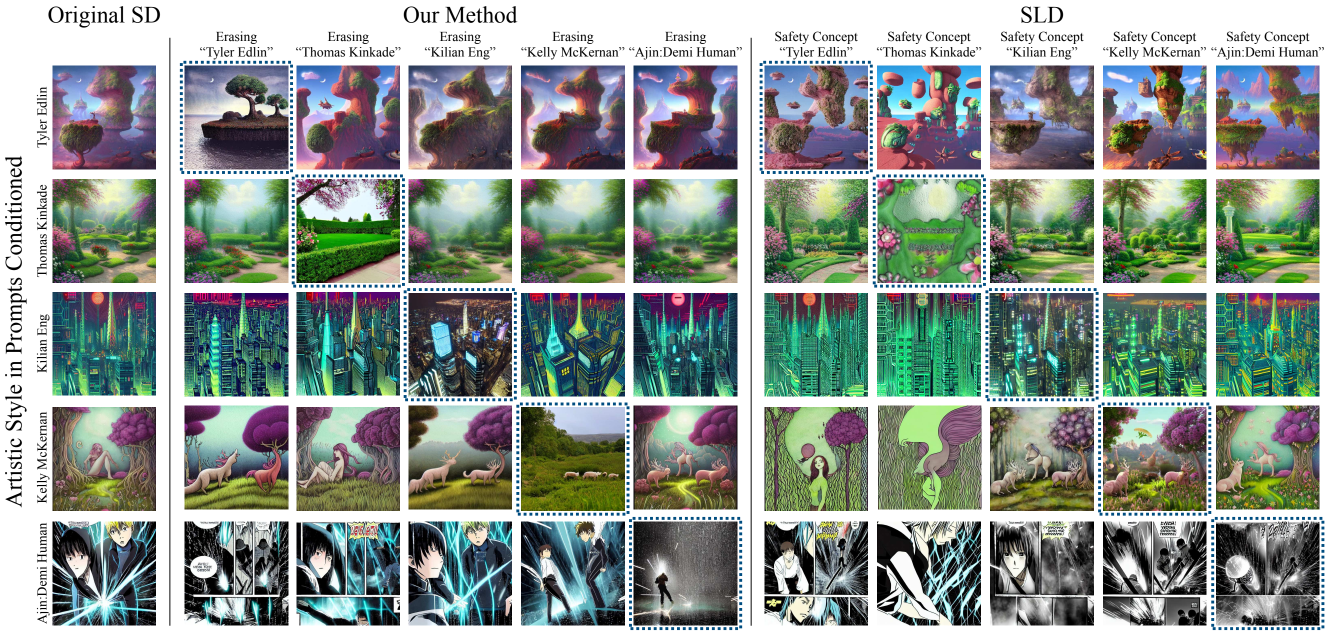

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Artistic Style in Prompts Conditioned
Tyler Edlin
Thomas Kinkade
Kilian Eng
Kelly McKernan
Ajin:Demi Human

Figure D.8: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

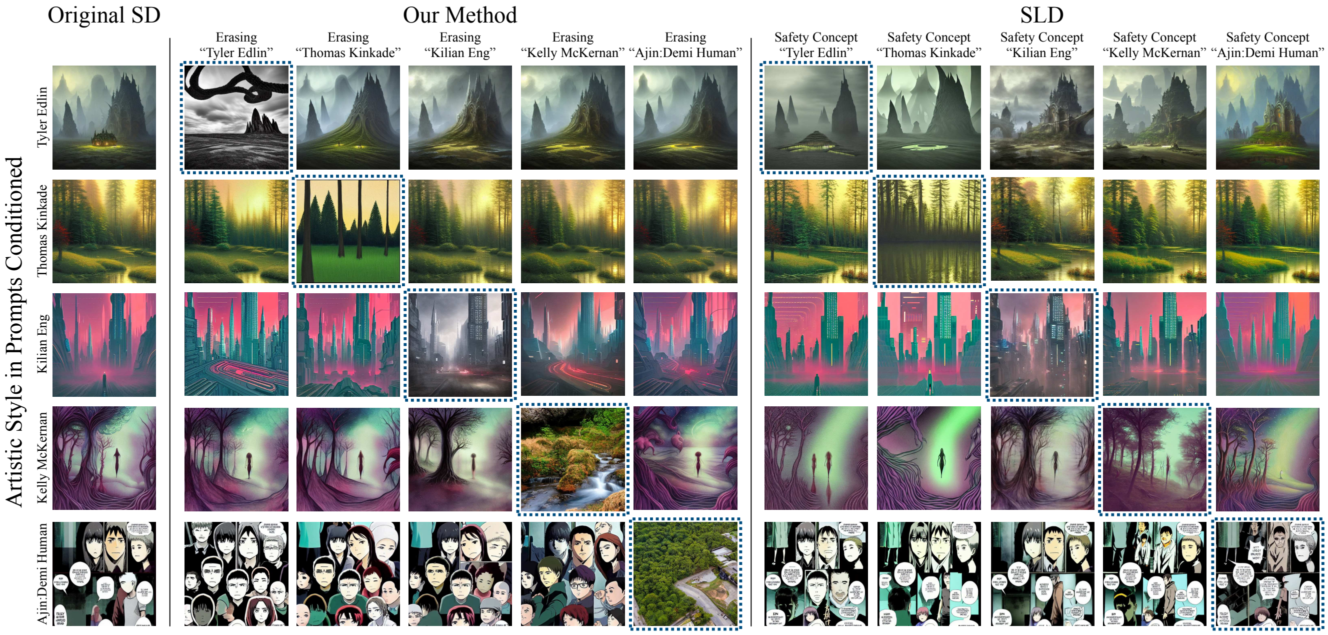

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Artistic Style in Prompts Conditioned
Tyler Edlin
Thomas Kinkade
Kilian Eng
Kelly McKernan
Ajin:Demi Human

Figure D.9: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

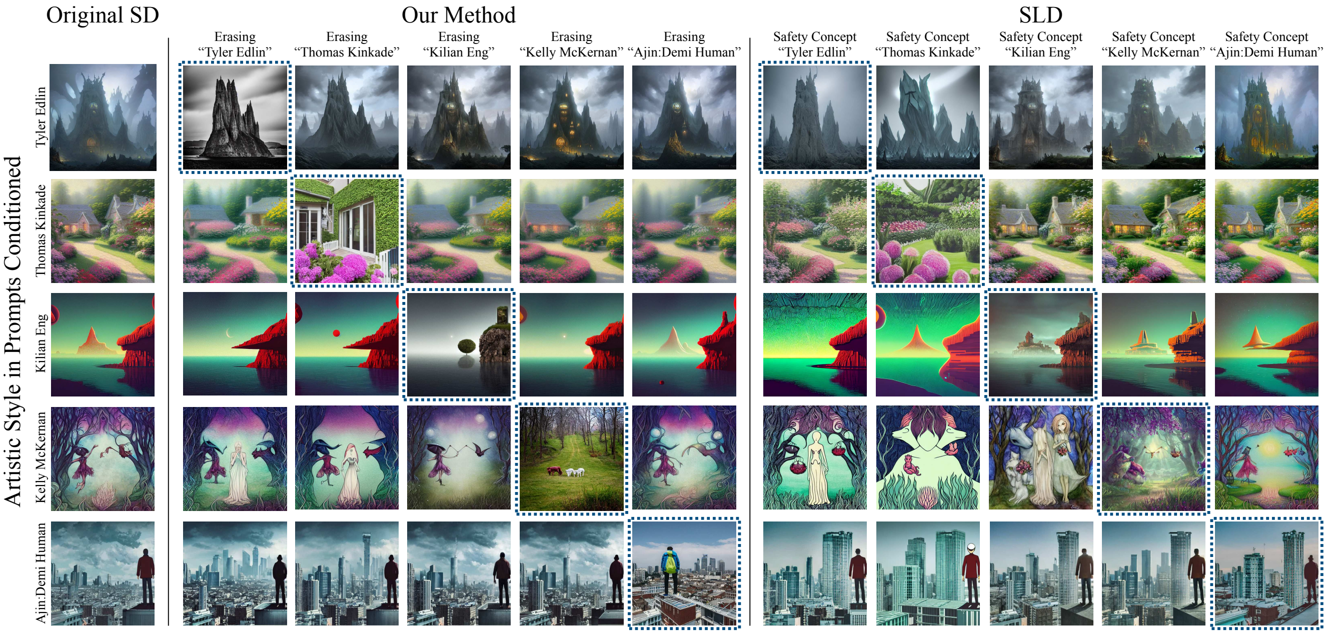

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Artistic Style in Prompts Conditioned
Tyler Edlin
Thomas Kinkade
Kilian Eng
Kelly McKernan
Ajin:Demi Human

Figure D.10: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

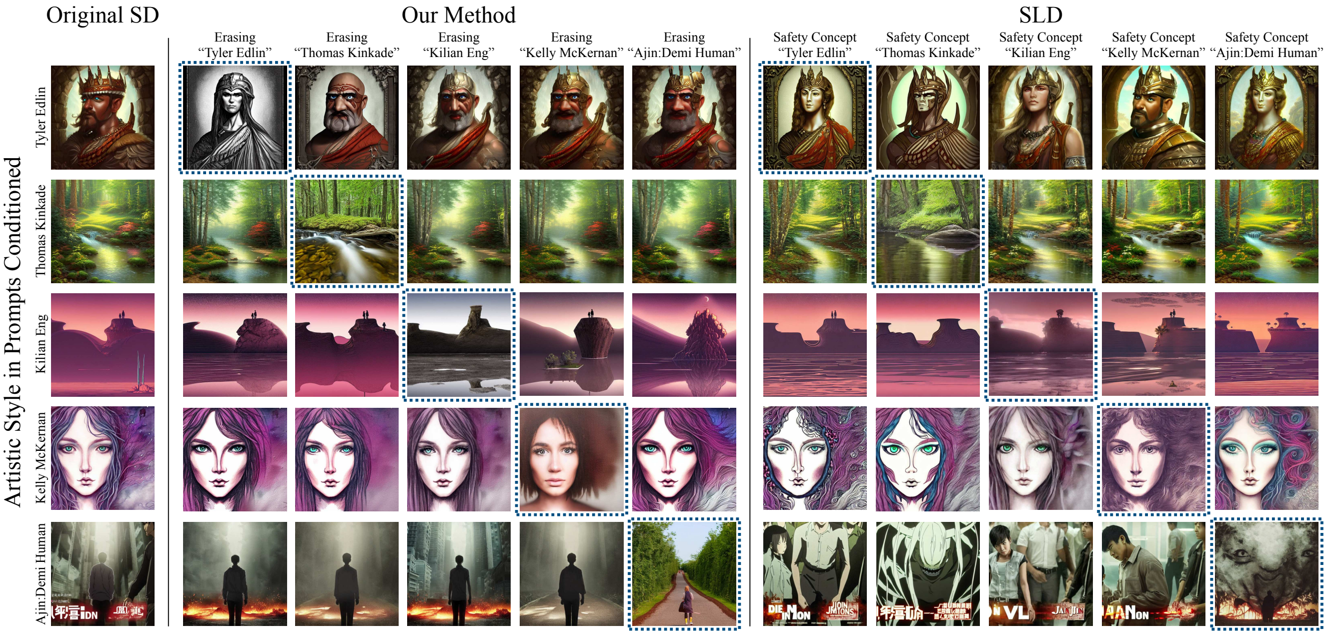

Original SD
Our Method
SLD
Erasing "Tyler Edlin"
Erasing "Thomas Kinkade"
Erasing "Kilian Eng"
Erasing "Kelly McKernan"
Erasing "Ajin:Demi Human"
Safety Concept "Tyler Edlin"
Safety Concept "Thomas Kinkade"
Safety Concept "Kilian Eng"
Safety Concept "Kelly McKernan"
Safety Concept "Ajin:Demi Human"
Artistic Style in Prompts Conditioned
Tyler Edlin
Thomas Kinkade
Kilian Eng
Kelly McKernan
Ajin:Demi Human

Figure D.11: Our method demonstrates a complete erasure of intended style and minimal interference with other styles. The blue dotted boxes show images with intended style erased. The off-diagonal images show the unintended interference.

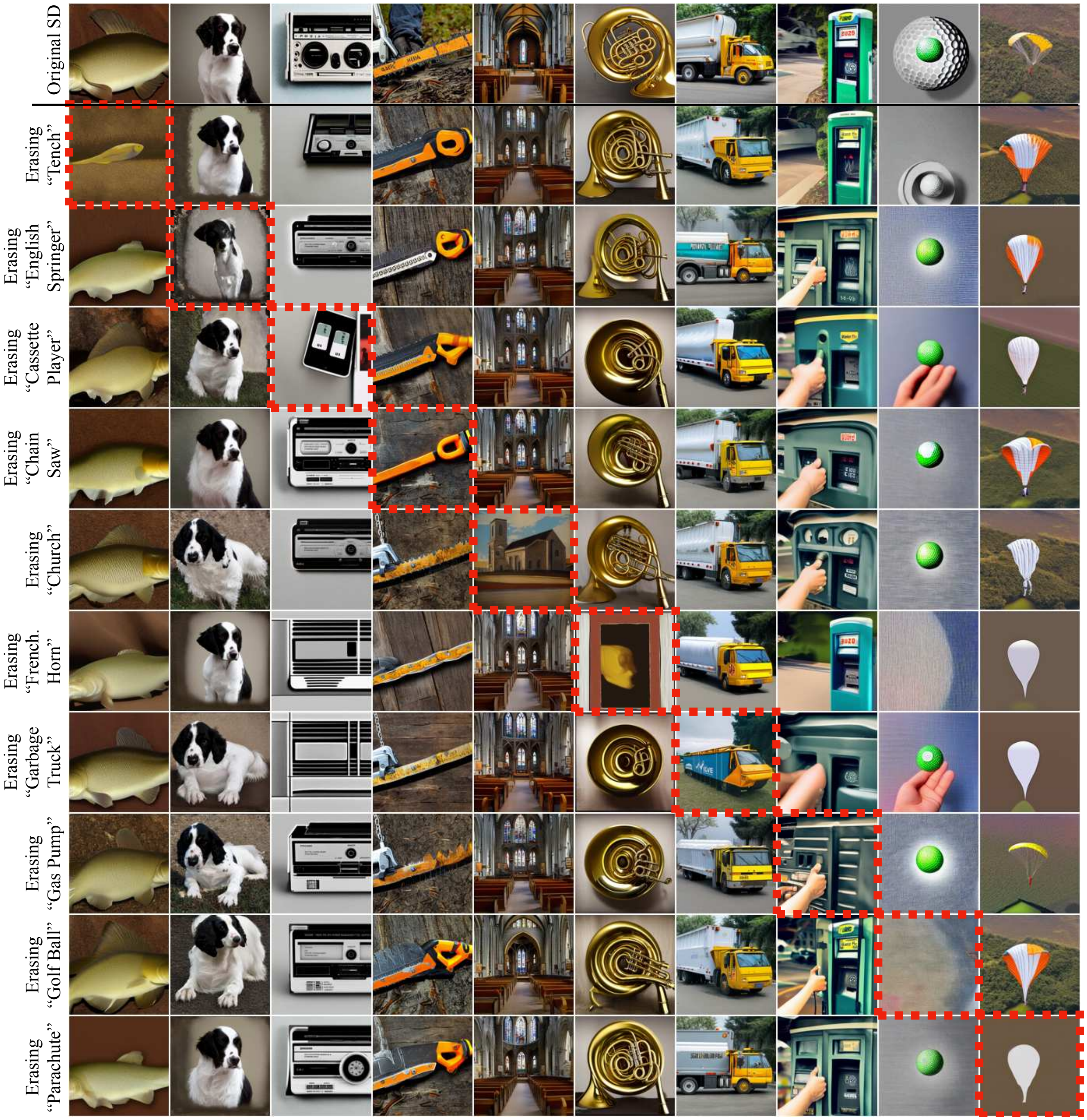


Figure D.12: Object removal in Stable Diffusion. The first row represents the original SD generations. From the later rows, the diagonal images represent the intended erasures while the off-diagonal images represent the interference.

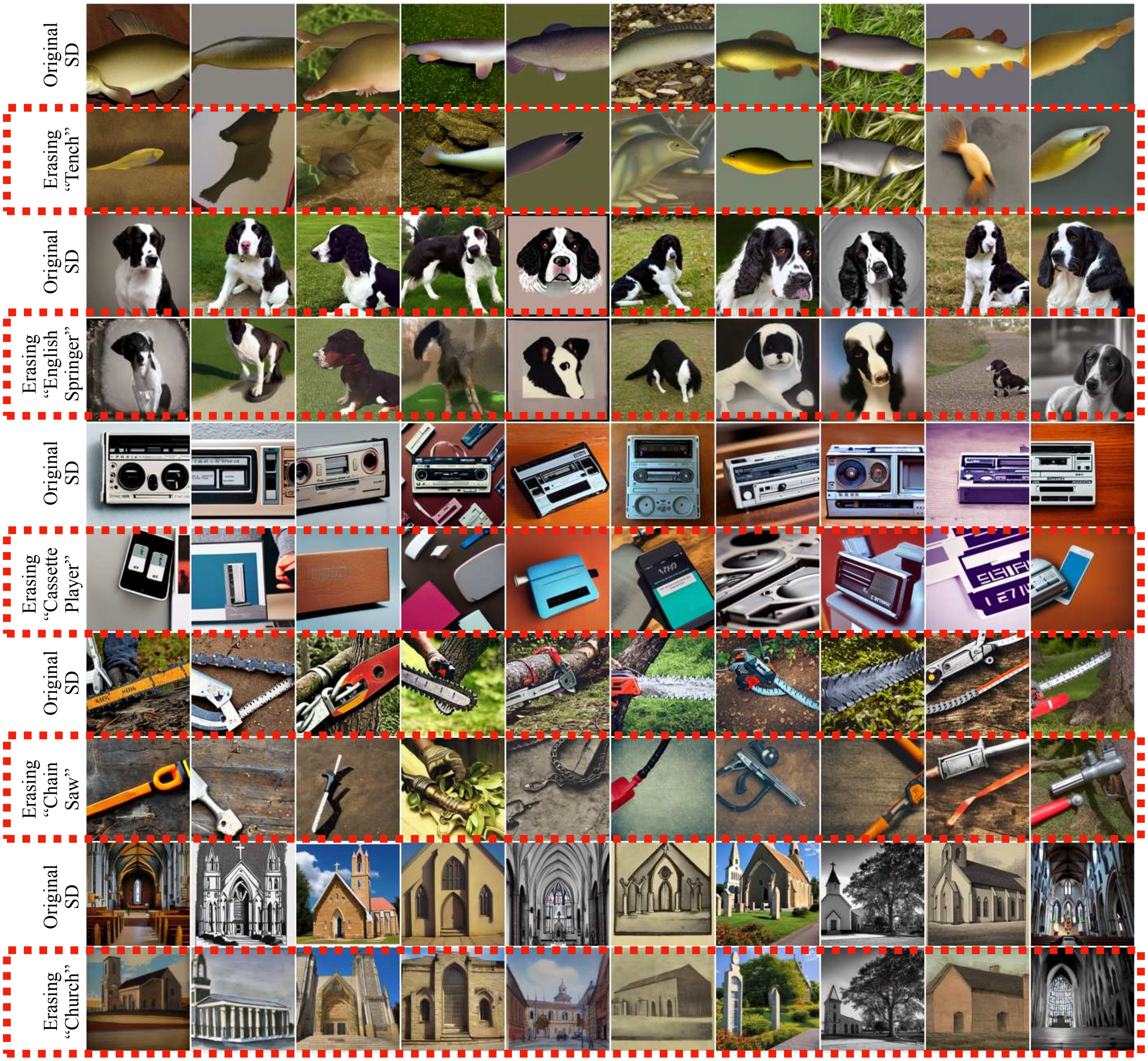


Figure D.13: We show the intended erasure of objects by our method (Part 1). The rows in red-dotted box represent erasure of an object while the row above each of the red boxes represent their corresponding original SD image using the same seed and prompts.

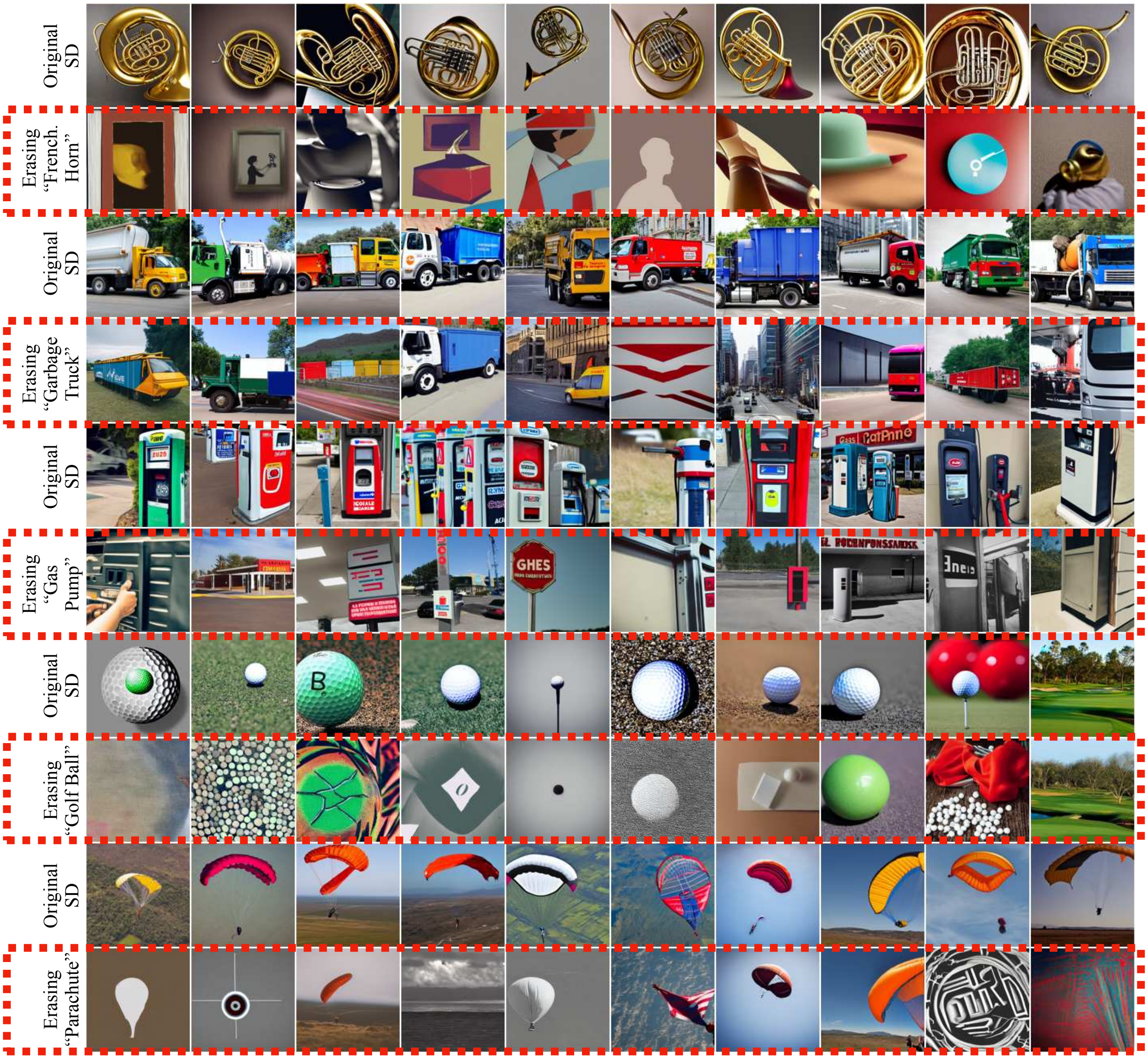


Figure D.14: We show the intended erasure of objects by our method (Part 2). The rows in red-dotted box represent erasure of an object while the row above each of the red boxes represent their corresponding original SD image using the same seed and prompts.