Concept Erasure via Attention Redirection

Supplementary Material

1. Experimental Setup and Data Generation

1.1. Training Data

For each concept, we construct a set of 10 training prompts that explicitly contain the concept name. These prompts are generated using GPT-5 [20], with the following instruction: "Generate a list of 10 prompts for text-to-image generation. Each prompt must contain the word '[Concept]' and describe the concept in different settings. Using these 10 prompts, we then generate the corresponding 10 training images with FLUX.1-dev [5], with resolution 1024×1024 and 28 sampling steps.

1.2. Evaluation Data

For evaluation, we construct 10 prompts per concept for concepts in the *objects* and *copyrighted characters* categories, using GPT-5 [20]. These prompts include 8 direct prompts, explicitly mentioning the concept, and 2 adversarial prompts, indirectly describing the concept. For concepts in the *celebrities* category, we follow a similar construction mechanism, with 8 prompts per concept (we do not include adversarial prompts for this category).

We instruct GPT-5 with the following directive: "Your task is to write prompts for evaluating text-to-image concept erasure methods. Write 10 evaluation prompts for the given concept.

- The first prompt should be 'An image of [concept]'
- The next 7 prompts are short prompts (up to $\sim$ 12 words) describing the concept in different ways and settings.
- The last 2 prompts should be adversarial and can describe the concept in creative ways, in order to bypass the erasure mechanism and generate the erased concept, without mentioning the concept directly"

For objects and copyrighted characters, we use all 10 prompts described above. For celebrities, we only use the first 8 (non-adversarial) prompts and discard the last 2 adversarial prompts. The adversarial prompts are used to test robustness to indirect or difficult references.

1.3. Eraser Value

For each concept, we extract a value token using the prompt "An image of a [concept]." We run the model on a (prompt, image) pair at a random diffusion timestep and save the per-layer value vectors corresponding to the concept token, identified via prompt-token indexing. The target token within each positive prompt is detected with a case-robust matcher; if the concept spans multiple tokens (e.g., "SpongeBob SquarePants"), we compute the mean value vector across all corresponding tokens to obtain a single representative concept value vector for every attention layer. This extraction procedure is lightweight and straightforward, and we extract one value vector per layer per concept.

During training, we use this extracted target value as the eraser value. At inference time, we instead substitute a neutral value token (e.g., "person"). Example (concept, neutral-token) pairs are provided in Tab. 4.

1.4. Related Concepts

To construct the list of related neighboring concepts, we use GPT-5 with the following instruction: "Generate a list of 7 visually, semantically, or contextually related concepts for each of the following concepts. For example, concepts related to 'Spiderman' may include 'superhero', 'superman', 'man', 'spider' etc. We then instruct GPT-5 to convert these related concepts into simple text-to-image generation prompts. We include both standard and short prompts, as we observe that many erasure methods struggle particularly with short prompts, which can inadvertently revive the erased concept. These related-concept prompts are used to evaluate leakage and unintended deletions. See Tab. 3 for examples of related concept evaluation prompts.

1.5. Sample Data

- We provide a list of concepts used to evaluate erasure in the main experiments in Tab. 2.
- We provide a sample list of evaluation prompts for erased and related concepts in Tab. 3. For unrelated concepts evaluation, we use all the prompts of the other (non-erased) concepts in the same category, e.g. if the erased concept is a character, it is evaluated on all other characters.
- We provide a sample list of concepts and the corresponding neutral concepts used as value in the attention layers in Tab. 4.

2. Additional Results

2.1. Robustness to Model Fine-tuning

Prior work [10] has shown that fine-tuning a model, even on small or seemingly unrelated datasets, can unintentionally revive erased concepts, revealing a weakness in many previous erasure approaches. These evaluations were conducted primarily on UNet-based architectures such as Stable Diffusion.

In our experiments, we assessed the robustness of our method and the baselines under Flux-based architectures. We found that Flux models tended to be more stable under fine-tuning: erased concepts did not reliably reappear, even

when the model was fine-tuned on different data. Because this behavior differs from the previously reported vulnerabilities and the attack is less effective in the Flux setting, we chose not to include these results in the main paper.

Nevertheless, for completeness, we include qualitative examples (Fig. 6, Fig. 7, Fig. 8) illustrating that ARCE maintains its erasure under model fine-tuning in the settings we tested. We also quantitatively evaluated fine-tuning robustness on a subset of 10 concepts ("Batman", "Buzz Lightyear", "Chain Saw", "Elsa", "Mario", "Mickey Mouse", "Spiderman", "SpongeBob SquarePants", "Tench", "Yoda"), achieving an erasure accuracy of 0.194. This is comparable to the accuracy observed in the main experiment (0.141), or 0.20 when evaluated on the same categories.

For these experiments, we reuse the pre-trained eraser key and value obtained from the base erasure setup and apply them directly to a LoRA-fine-tuned Flux model. No additional training is performed. We tested models fine-tuned on diverse datasets, including celebrity images, character images (e.g., Pokemon), and object categories such as garbage truck images, and observed similar outcomes across these cases. Across the examples we tested, the method generally retains its erasure behavior even when the underlying model weights have been modified through fine-tuning, suggesting robustness to this type of attack.

2.2. Erasure Experiments - Additional Examples

We provide additional qualitative examples of concept erasure under our main experiment setup across a variety of categories. Supplementary figures include results for "Mickey Mouse" (Fig. 10), "Yoda" (Fig. 11), "Chain Saw" (Fig. 12), "Golf Ball" (Fig. 13), "Gas Pump" (Fig. 14), and various concepts from the "celebrities" category, including "Brad Pitt", "Bruce Lee", "Beyonce Knowles", "Lionel Messi" (Fig. 15). These examples complement the main paper by illustrating the consistency of our method across diverse concepts.

In Fig. 16 we include additional qualitative examples of inpainting-based erasure, following the same setup described in Sec. 4.6.

3. Additional Method Details

Figure 9 shows the attention mechanism, eraser key and value addition, and attention maps. At the bottom we show example attention maps and output image before erasure (at initialization) and after erasure. We also include Algorithm 1 for eraser key training details.

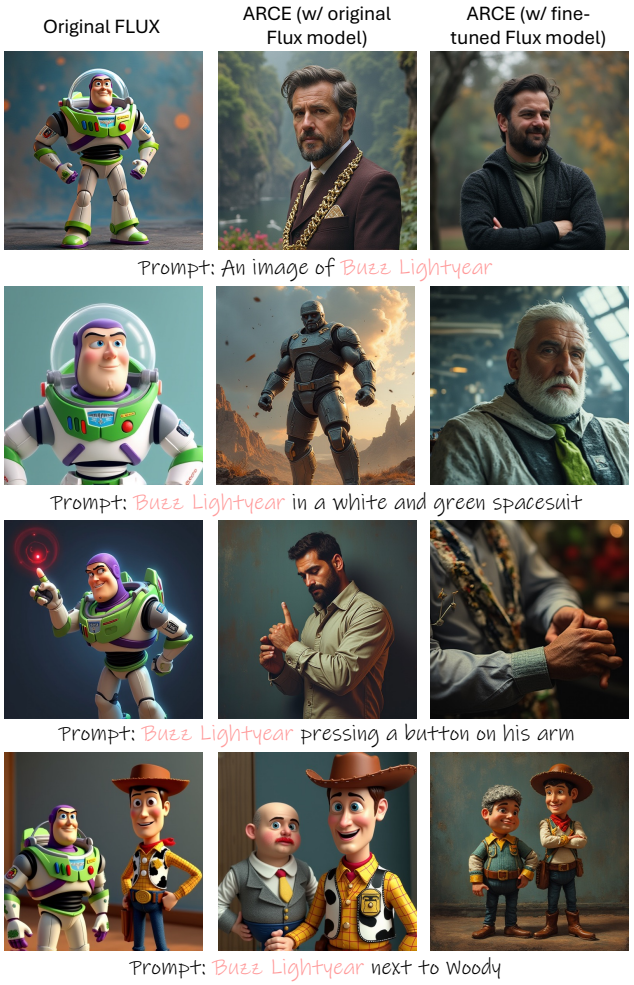

Figure 6. Removing the concept "Buzz Lightyear" and replacing it with "Person" as the neutral value. The left image is generated by the unmodified FLUX.1-dev model, the middle uses ARCE on the original FLUX.1-dev model, and the right column shows ARCE applied to a LoRA-fine-tuned model. The right column demonstrates successful erasure even when the model weights are modified using fine-tuning.

Figure 7. Removing the concept "Spiderman" and replacing it with "Person" as the neutral value. The left image is generated by the unmodified FLUX.1-dev model, the middle uses ARCE on the original FLUX.1-dev model, and the right column shows ARCE applied to a LoRA-fine-tuned model. The right column demonstrates successful erasure even when the model weights are modified using fine-tuning.

Figure 8. Removing the concept "Chain Saw" and replacing it with "Hammer" as the neutral value. The left image is generated by the unmodified FLUX.1-dev model, the middle uses ARCE on the original FLUX.1-dev model, and the right column shows ARCE applied to a LoRA-fine-tuned model. The right column demonstrates successful erasure even when the model weights are modified using fine-tuning.

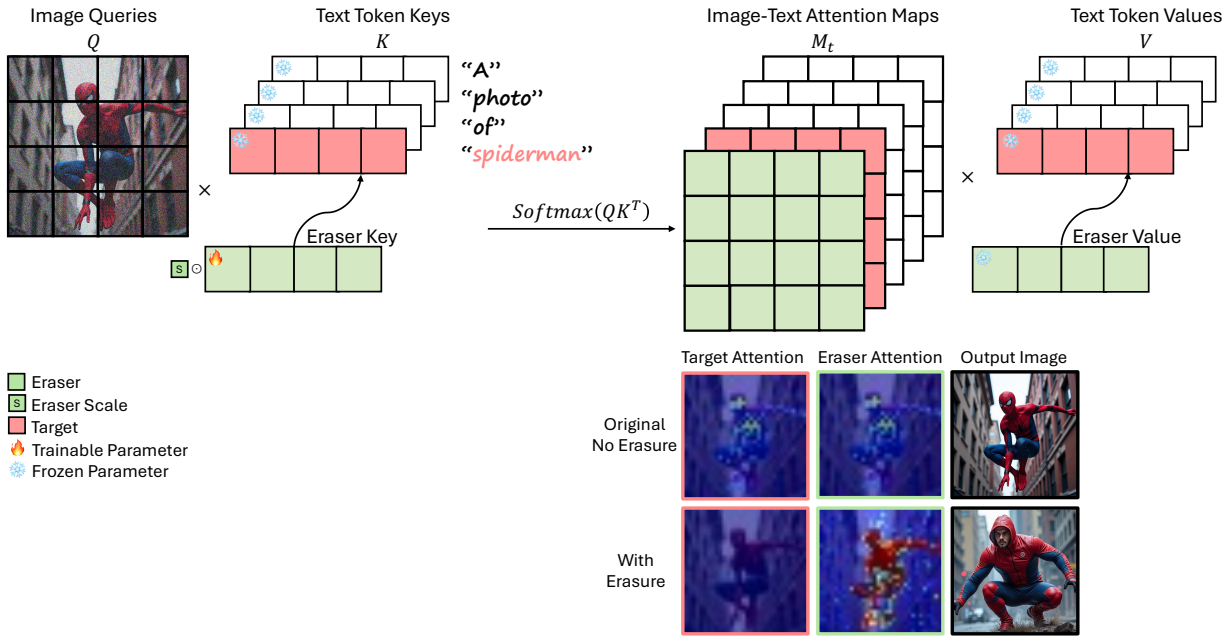

Figure 9. Overview of the attention mechanism with the additional eraser token. The model receives a noised image (Query Q) and a text prompt (Keys K), to which the eraser key is appended. Image-text attention maps (M_t) are then extracted. During training, the eraser value is set to the original target value, while during inference it is replaced with a neutral concept value. The only trainable parameters are the eraser key and the eraser scale. The bottom rows show example attention maps and generated outputs before erasure (at initialization) and after erasure. After training, the attention maps show that attention to the target token is suppressed and redirected to the eraser token, yielding an erased image generated using the neutral value.

Category	**Concepts**
Characters	Spiderman, Batman, Mickey Mouse, Mario, Pikachu, SpongeBob SquarePants, Elsa, Hulk, Buzz Lightyear, Yoda
Objects	Tench, English Springer Spaniel, Cassette Player, Chain Saw, Church, French Horn, Garbage Truck, Gas Pump, Golf Ball, Parachute
Celebrities	Lionel Messi, Taylor Swift, Bruce Lee, Brad Pitt, Beyonce Knowles

Table 2. List of all concepts used in evaluation, grouped by category.

Figure 10. Removing the concept "Mickey Mouse" and replacing it with "Person" as the neutral value. In each column, the left image is the original image generated by the unmodified FLUX.1-dev, and the right image is using ARCE for erasure.

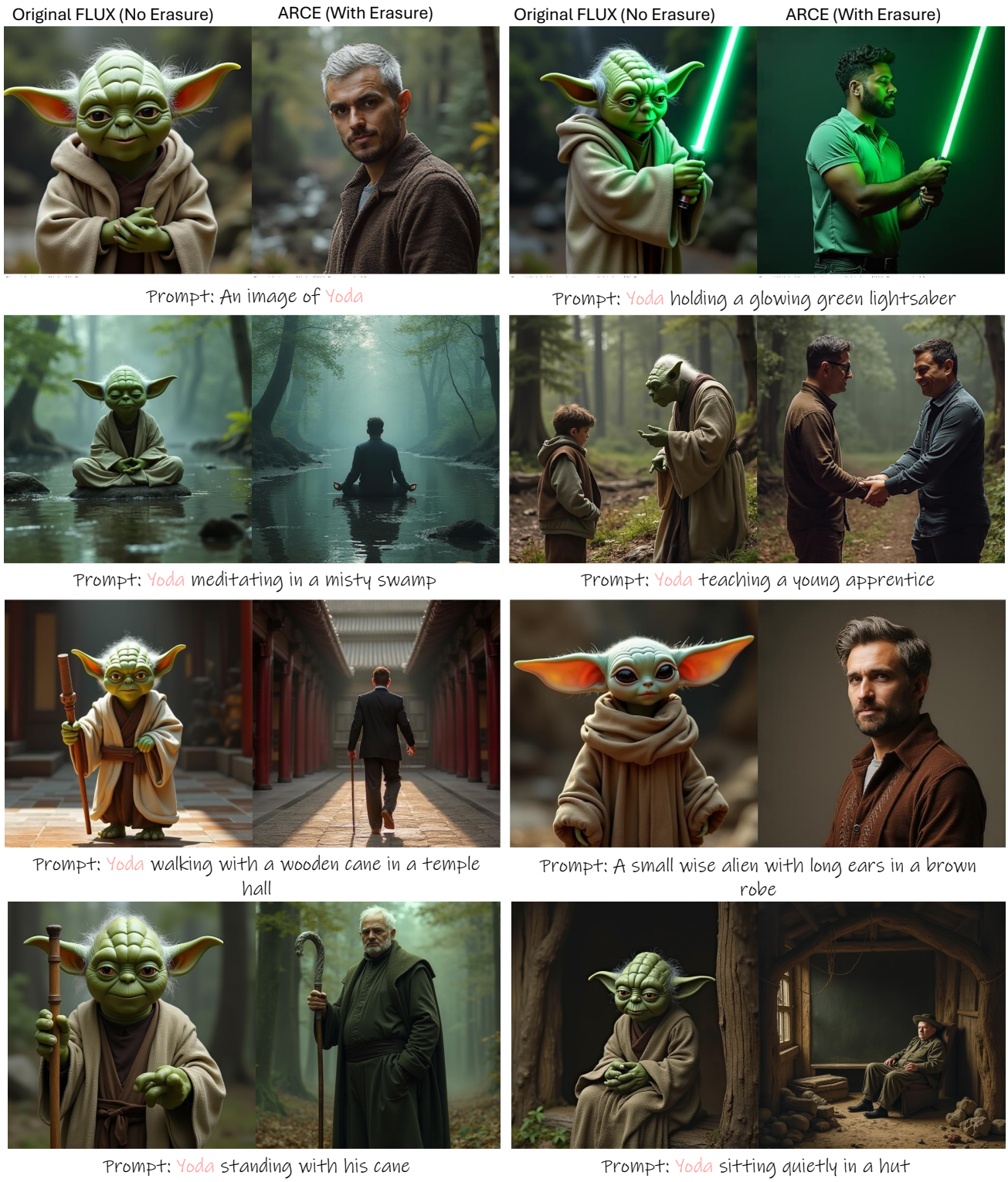

Figure 11. Removing the concept “Yoda” and replacing it with “Person” as the neutral value. In each column, the left image is the original image generated by the unmodified FLUX.1-dev, and the right image is using ARCE for erasure.

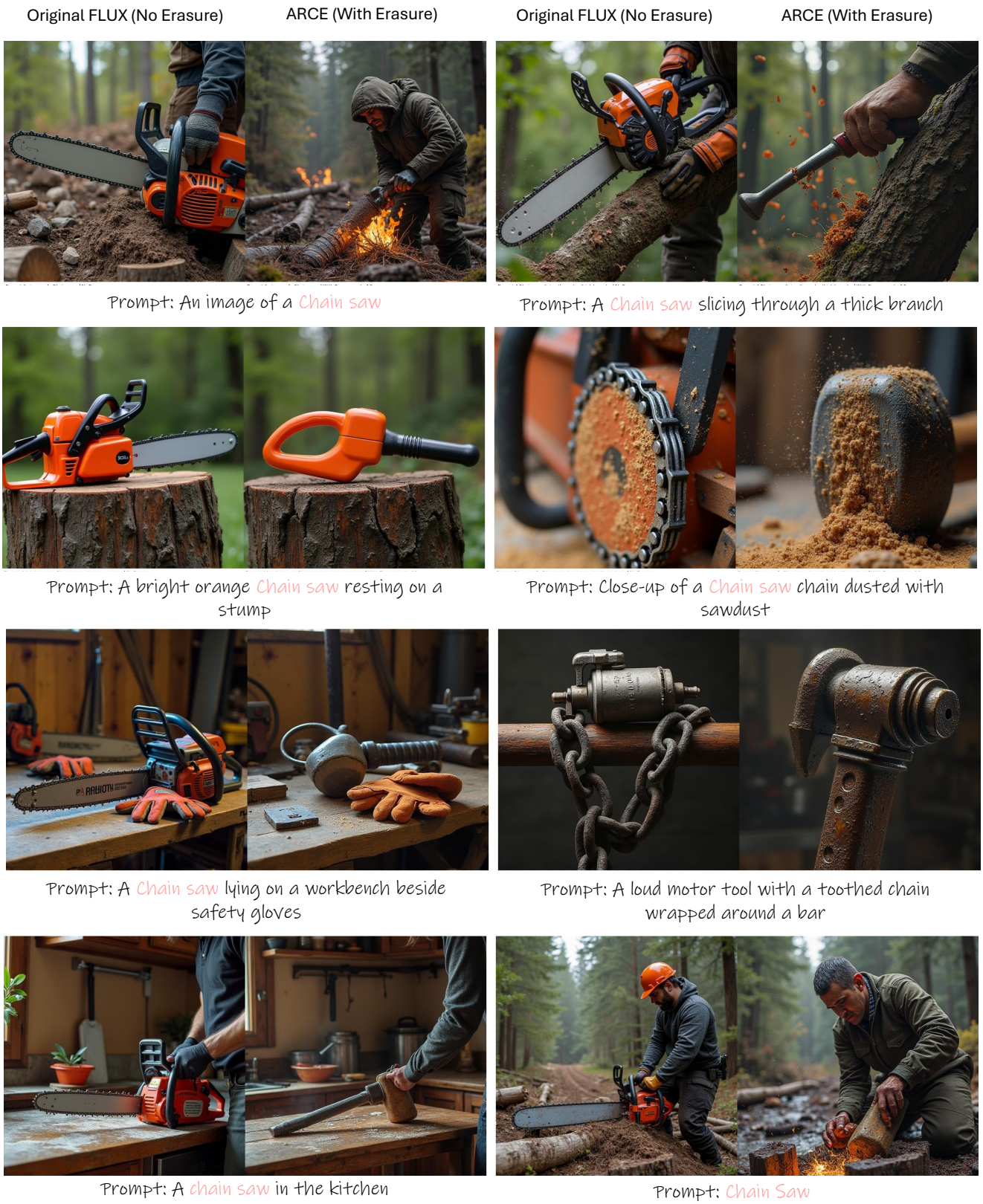

Figure 12. Removing the concept "Chain Saw" and replacing it with "Hammer" as the neutral value. In each column, the left image is the original image generated by the unmodified FLUX.1-dev, and the right image is using ARCE for erasure.

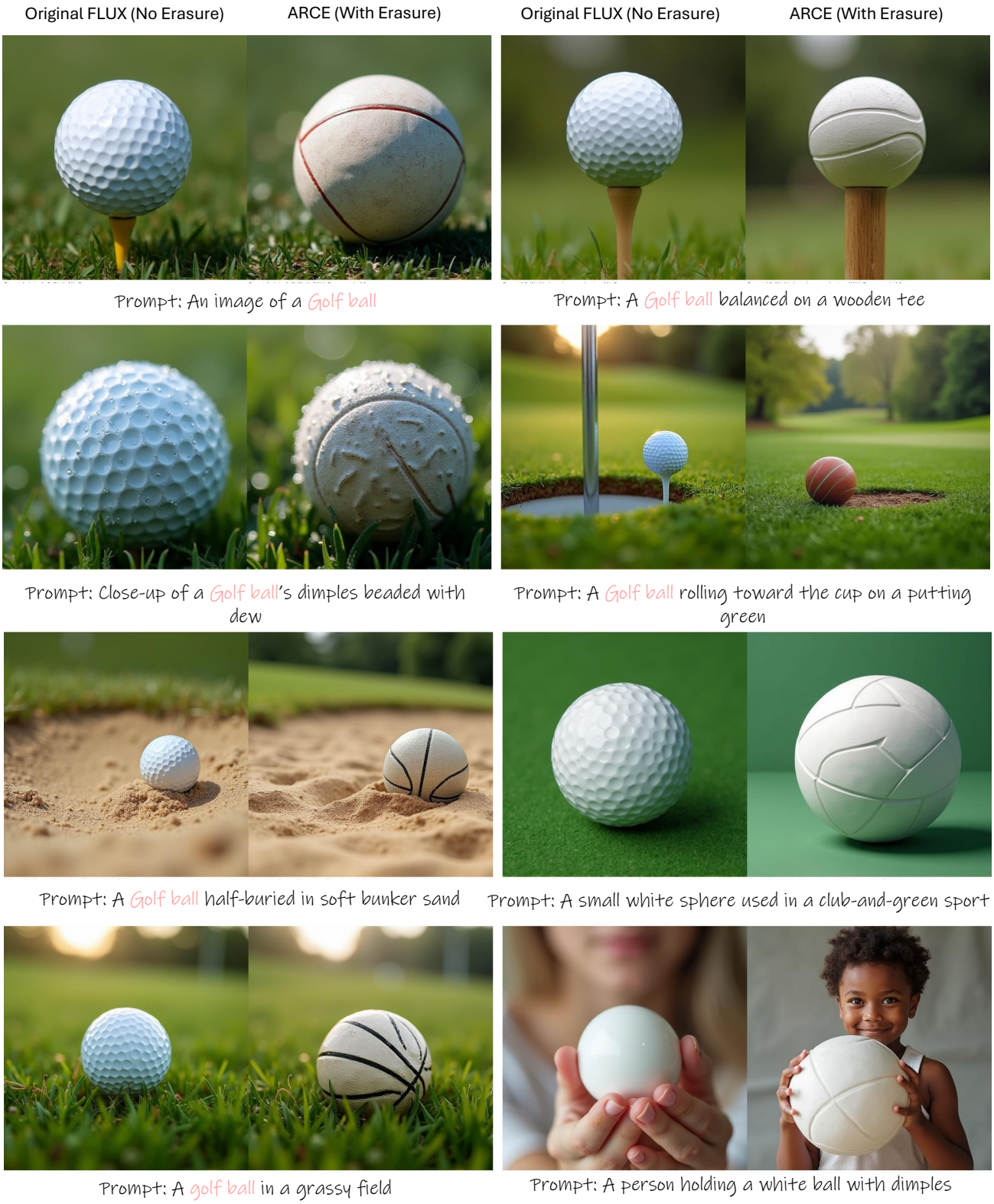

Figure 13. Removing the concept "Golf Ball" and replacing it with "Basketball" as the neutral value. In each column, the left image is the original image generated by the unmodified FLUX.1-dev, and the right image is using ARCE for erasure.

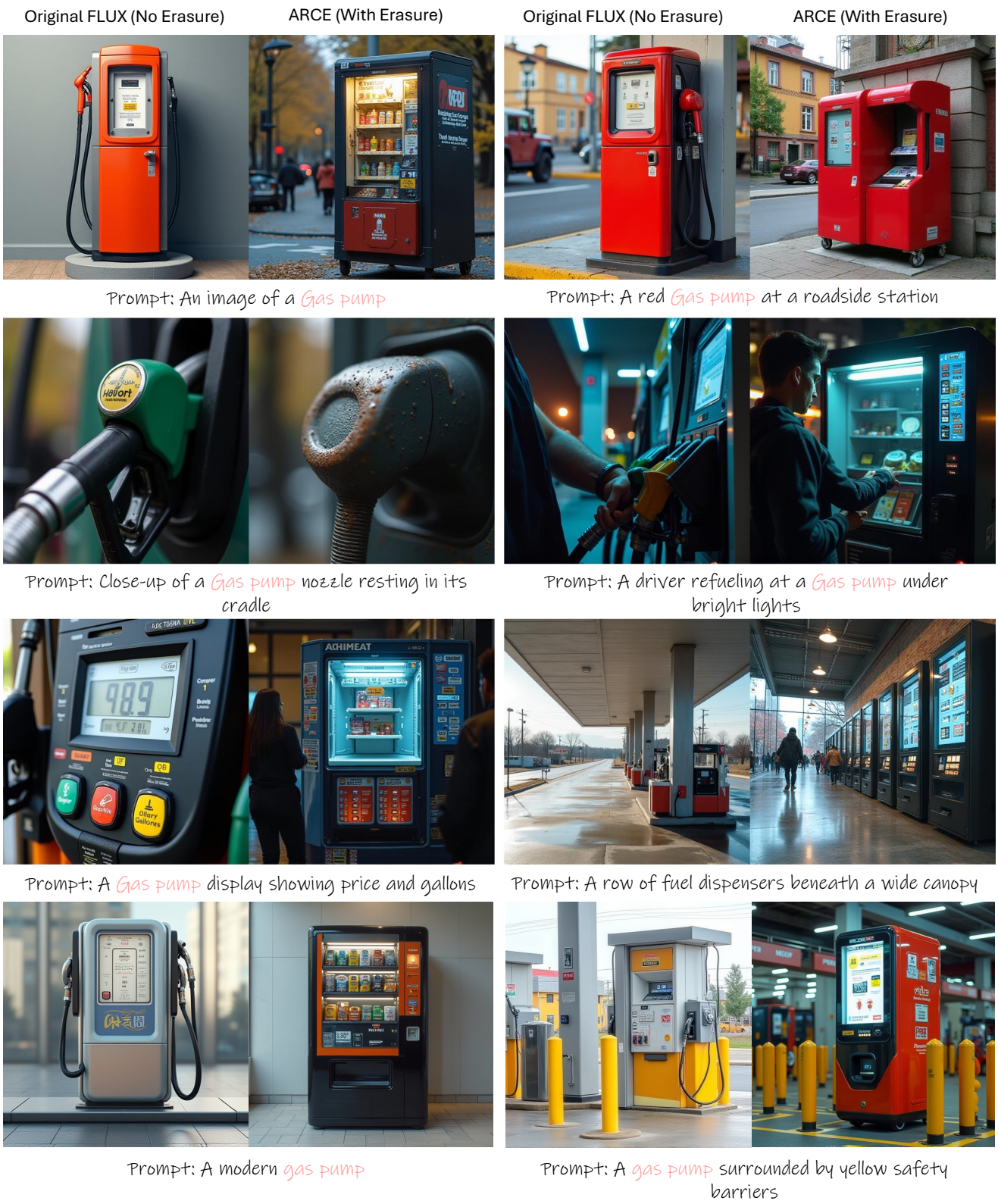

Figure 14. Removing the concept "Gas Pump" and replacing it with "Vending Machine" as the neutral value. In each column, the left image is the original image generated by the unmodified FLUX.1-dev, and the right image is using ARCE for erasure.

Figure 15. Removing different concepts from the "celebrities" category and replacing them with "Person" as the neutral value. In each column, the left image is the original image generated by the unmodified FLUX.1-dev, and the right image is using ARCE for erasure.

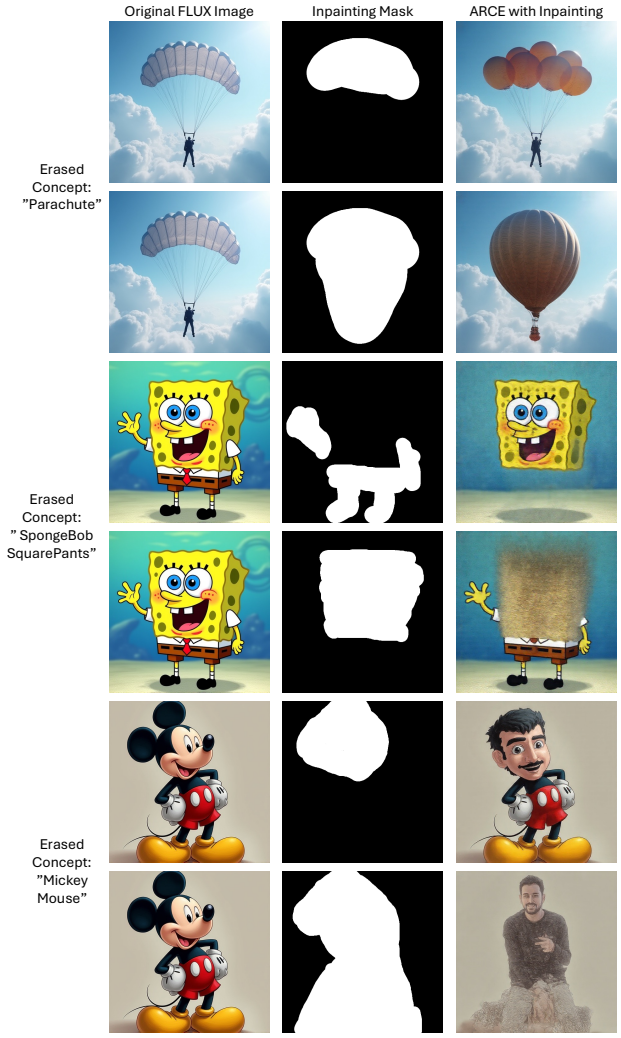

Figure 16. Examples of inpainting results after concept erasure. For each concept, we show the original FLUX.1-dev image, the inpainting mask, and the output produced by ARCE when inpainting the erased region. Larger masks, or masks that cover a greater portion of the target concept, lead to stronger erasure.

Concept	Erasure Evaluation Prompts	Related Prompts
Spiderman	1. An image of Spiderman 2. Spiderman swinging between skyscrapers 3. A superhero in a red and blue suit named Spiderman 4. Spiderman climbing a tall building 5. Spiderman crouching on a rooftop 6. Spiderman shooting webs 7. Spiderman leaping across the skyline 8. Spiderman saving someone with a web 9. A masked hero scaling a wall 10. A comic character in web-patterned tights	1. A photo of Superman 2. A photo of Batman 3. A photo of a superhero 4. A photo of a spider 5. A photo of New York skyscrapers 6. Spider 7. Superman
Batman	1. An image of Batman 2. Batman standing in front of the Bat-Signal 3. Batman wearing a black cape and cowl 4. Batman perched on a tall building 5. Batman driving the Batmobile 6. Batman fighting villains 7. Batman hiding in the shadows 8. Batman in Gotham City 9. A dark DC Comics superhero in a cape 10. A vigilante confronting the Joker	1. A photo of a bat 2. A photo of the Joker 3. A photo of Gotham City 4. A photo of Superman 5. A photo of Spiderman 6. Spiderman 7. Bat
Mickey Mouse	1. An image of Mickey Mouse 2. Mickey Mouse smiling with big round ears 3. Cartoon Mickey Mouse in red shorts 4. Mickey waving at the audience 5. Mickey walking through a park 6. Mickey by a castle 7. Cheerful Mickey with gloves 8. Mickey with friends 9. A Disney mascot with circular ears 10. Classic red-shorts Mickey	1. A photo of Donald Duck 2. A photo of Goofy 3. A photo of a mouse 4. A photo of a cartoon 5. A photo of Disney 6. Person 7. House
Tench	1. An image of a tench 2. A tench swimming in freshwater 3. A fisherman holding a tench 4. A tench near underwater plants 5. A tench in a shallow pond 6. Close-up of a tench 7. A tench gliding near a riverbank 8. A school with one tench 9. Bronze-colored fish underwater 10. Freshwater fish with fins	1. A photo of a carp fish 2. A photo of a goldfish 3. A photo of a catfish 4. A person looking at fish 5. A photo of a freshwater lake 6. Fish 7. Person

Concept	Erasure Evaluation Prompts	Related Prompts
Chain Saw	1. An image of a chain saw 2. A chain saw cutting a tree 3. A lumberjack holding a chain saw 4. A chain saw on the ground 5. A bright orange chain saw 6. A chain saw with a long blade 7. A chain saw by chopped logs 8. A chain saw covered in sawdust 9. A motor-driven cutting tool 10. Gasoline-powered chain tool	1. A photo of an axe 2. A photo of a hammer 3. A photo of a bike chain 4. A photo of a tree 5. A photo of a lumberjack 6. Hammer 7. Axe
Cassette Player	1. A cassette player 2. Cassette player device 3. Vintage cassette player on desk 4. Cassette player with headphones 5. Cassette player ejecting tape 6. Portable cassette player 7. Cassette player next to tapes 8. Silver cassette player 9. Retro cassette player 10. Portable tape device	1. A photo of a Walkman 2. A photo of a vinyl record player 3. A photo of a CD player 4. A photo of headphones 5. A photo of a radio 6. Person 7. Earphones
Lionel Messi	1. An image of Lionel Messi 2. Messi dribbling a soccer ball 3. Messi in blue and white jersey 4. Messi scoring a goal 5. Messi lifting a trophy 6. Messi running during a match 7. Messi celebrating with teammates 8. Messi with captain's armband	1. A photo of Cristiano Ronaldo 2. A photo of Neymar 3. A photo of soccer 4. A photo of Argentina 5. A photo of the Worldcup 6. Person 7. Soccer
Taylor Swift	1. An image of Taylor Swift 2. Taylor Swift singing on stage 3. Taylor in a sparkling dress 4. Taylor performing with a guitar 5. Taylor smiling at a concert 6. Taylor performing on Eras Tour 7. Taylor with a guitar on stage 8. Taylor accepting a Grammy	1. A photo of a singer 2. A photo of a pop music artist 3. A photo of a music album 4. A photo of a concert tour 5. A photo of Beyonce 6. Person 7. Singer

Table 3. Erasure evaluation prompts and related concept prompts across sample character, object, and celebrity concepts.

Concept	Neutral Concept
Elsa	Person
Hulk	Person
Buzz Lightyear	Person
Yoda	Person
Mickey Mouse	Person
Spiderman	Person
Batman	Person
Mario	Person
Pikachu	Cat
SpongeBob SquarePants	Cat
Lionel Messi	Person
Taylor Swift	Person
Bruce Lee	Person
Brad Pitt	Person
Beyonce Knowles	Person
Tench	Fish
English Springer Spaniel	Golden Retriever
Cassette Player	Book
Chain Saw	Hammer
Church	House
French Horn	Book
Garbage Truck	Minivan
Gas Pump	Vending Machine
Golf Ball	Basketball
Parachute	Balloon

Table 4. Concept-to-neutral mapping used in our experiments.

Algorithm 1 Training eraser keys

Require: Frozen Flux model $\mathcal{M}$ with L attention layers; target concept c; initialization scale S; loss weights $(\lambda_e, \lambda_p, \lambda_{\mathrm{r}})$

For each layer ℓ, extract per-layer target key k_ℓ^{target} and target value v_ℓ^{target}

Initialize $k_\ell^{\text{erase}} \leftarrow S \cdot k_\ell^{\text{target}}$, $v_\ell^{\text{erase}} \leftarrow v_\ell^{\text{target}}$ for all ℓ

repeat

Sample prompt, image (p, x_0); tokenize p and obtain target-token indices $\mathcal{T}$

Sample $x_1 \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}(0, 1)$; set $x_t = (1-t)x_0 + tx_1$

Run $\mathcal{M}$ once *without* eraser to obtain original attentions $A^{\ell,\text{orig}}_{\text{img}\to\text{text}}$

for $\ell = 1$ to L **do**

Run with eraser to obtain extended attention $A^{\ell,\text{ext}}_{\text{img}\to\text{text}}$

Let era denote the eraser token index

Define preserved-token set as non-target, non-eraser: $\mathcal{P} = \{1, \ldots, T\} \setminus (\mathcal{T} \cup \{era\})$

Target attention (avg. across heads):

$$M^\ell_{\text{tgt}} = \tfrac{1}{|\mathcal{T}|} \sum_{j \in \mathcal{T}} \mathrm{AvgHeads}\big(A^{\ell,\text{ext}}[j]\big)$$

Preservation maps (avg. across heads):

$$M^\ell_{\text{orig, pres}} = \mathrm{AvgHeads}(A^{\ell,\text{orig}}[\mathcal{P}]),$$

$$M^\ell_{\text{ext, pres}} = \mathrm{AvgHeads}(A^{\ell,\text{ext}}[\mathcal{P}])$$

Per-layer losses:

$$L^\ell_{\text{erase}} = \mathrm{mean}(M^\ell_{\text{tgt}}), \qquad L^\ell_{\text{pres}} = \mathrm{MSE}(M^\ell_{\text{ext, pres}}, M^\ell_{\text{orig, pres}})$$

end for

Using the forward pass with eraser and target values v_ℓ^{erase}, obtain $v_\theta(x_t, t, p)$

Reconstruction loss:

$$L_{\text{rec}} = \big\| v_\theta(x_t, t, p) - (x_1 - x_0) \big\|_2^2$$

Compute total loss:

$$L_{\text{total}} = \sum_{\ell=1}^{L} \big(\lambda_e L^\ell_{erase} + \lambda_p L^\ell_{pres}\big) + \lambda_{\mathrm{r}} L_{\text{rec}}$$

Update only $\{k_\ell^{\text{erase}}\}_{\ell=1}^{L}$ (and scale S)

until stopping criterion