Appendix Catalogue

A Proof for Theorems **13**
- A.1. Additional Notations . . . 13
- A.2. Proof for Theorem 3.1 . . . 13
- A.3. Proof for Theorem 4.2 . . . 14
- A.4. Proof for Theorem 4.3 . . . 15

B Detailed Experiment Settings **15**
- B.1. Object Erasing . . . 15
- B.2. Evaluation Protocol for Object Erasure . . . 16
- B.3. Artist Style Erasure . . . 16

C Imperfection of Preservation **16**
- C.1. The Impacts of Positional Embedding . . . 16
- C.2. Self-Attention in Encoder . . . 17
- C.3. Why Editing on Embedding Layers Produces Better Preservation. . . . 17

D Why UCE Underperforms on *Cassette Player* and *Golf Ball* **17**
- D.1. Last Word Erasing . . . 17
- D.2. DP Still Achieves Better Preservation . . . 18

E Other Metrics for Image Assessment **19**

F. Additional Experiments on SD 1.5 **20**

G Visualization for Stable Diffusion 1.4 **21**

H Visualization for FLUX **21**

I. Ablation Studies **23**
- I.1 . Ablation on the First Projection . . . 23
- I.2 . Ablation on the Second Projection . . . 23

J. Generation on Other Classes **23**

A. Proof for Theorems

A.1. Additional Notations

Notations Let $W_0 \in \mathbb{R}^{p\times n}$ be the pretrained linear map, and $W = W_0 + \Delta W$ the updated map. Let $C_{\text{tgt}} = [c_1, \dots, c_T] \in \mathbb{R}^{n\times T}$ be target (to erase) embeddings, and $C_{\text{pres}} \in \mathbb{R}^{n\times m}$ the preserved set with $\text{rank}(C_{\text{pres}}) = r$. Let $C_{\text{pres}} = U_1 \Sigma V^\top$ be a thin SVD with left singular vectors $U_1 = [u_1, \dots, u_r] \in \mathbb{R}^{n\times r}$ ordered by singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Fix $k \in \{0, 1, \dots, r\}$ and define

$$U_k := [u_1, \dots, u_k] \in \mathbb{R}^{n\times k}, \quad U_{\text{tail}} := [u_{k+1}, \dots, u_r] \in \mathbb{R}^{n\times(r-k)}.$$

Let $U_{\text{out}} \in \mathbb{R}^{n\times(n-r)}$ be an orthonormal basis of the left nullspace of C_{pres}. Set the $(n-k)$-dimensional orthogonal complement of $\text{span}(U_k)$ as

$$U_{2,k} := \begin{bmatrix} U_{\text{tail}} & U_{\text{out}} \end{bmatrix} \in \mathbb{R}^{n\times(n-k)}, \qquad P_k := U_k U_k^\top, \quad I - P_k = U_{2,k} U_{2,k}^\top.$$

We parameterize the update by

$$\Delta W \;=\; Z\, U_{2,k}^\top, \qquad Z \in \mathbb{R}^{p\times(n-k)} \tag{P}$$

which enforces $\Delta W\, U_k = 0$ and thus *exactly preserves* the top-k preserve directions.

Let the *safe* proxy be $c_i^\star = \Pi_S c_i := S(S^\top S)^+ S^\top c_i$. Define

$$C_\perp^{(k)} := U_{2,k}^\top C_{\text{tgt}} \in \mathbb{R}^{(n-k)\times T}, \qquad B := W_0\big(C_{\text{tgt}}^\star - C_{\text{tgt}}\big) \in \mathbb{R}^{p\times T}.$$

The second projection reduces to a matrix least-squares problem

$$\min_Z \; \|Z\, C_\perp^{(k)} - B\|_F^2,$$

whose minimum-norm closed form is

$$Z^\star \;=\; B\,\big(C_\perp^{(k)}\big)^\top \Big(C_\perp^{(k)}\big(C_\perp^{(k)}\big)^\top\Big)^+, \qquad \Delta W^\star = Z^\star U_{2,k}^\top. \tag{CF}$$

A.2. Proof for Theorem 3.1

Proof. From the closed-form solution of UCE, the weight update is given by

$$\Delta W := W_{\text{uce}} - W_0 = (v^* - W_0 c)\, c^\top N^{-1},$$

where $N = cc^\top + C_{\text{pres}} C_{\text{pres}}^\top$. This expression shows that ΔW is a rank-one update: it modifies the weights in the direction of the residual $(v^* - W_0 c)$, scaled by a transformed version of the target vector c through N^{-1}.

To analyze how ΔW affects different representations, consider its action on the target vector c and a preserve vector p:

$$\Delta c = \Delta W c = (v^* - W_0 c)\,(c^\top N^{-1} c),$$
$$\Delta p = \Delta W p = (v^* - W_0 c)\,(c^\top N^{-1} p).$$

Both perturbations are proportional to the same direction $(v^* - W_0 c)$, but differ in magnitude depending on how p aligns with c in the metric defined by N^{-1}.

The coefficients $c^\top N^{-1} p$ and $c^\top N^{-1} c$ can be rewritten as inner products in a transformed space:

$$c^\top N^{-1} p = \langle N^{-1/2} c, N^{-1/2} p\rangle, \qquad c^\top N^{-1} c = \langle N^{-1/2} c, N^{-1/2} c\rangle.$$

This formulation highlights that the relative effect of ΔW on p depends on the correlation between $N^{-1/2} c$ and $N^{-1/2} p$. If p is well aligned with c under this transformation, it will inevitably experience a nontrivial perturbation when c is edited.

Since both Δp and Δc are parallel to $(v^* - W_0 c)$, their Euclidean norms differ only by the magnitude of the scalar coefficients:

$$\|\Delta p\|_2 = |c^\top N^{-1} p|\, \|v^* - W_0 c\|_2, \quad \|\Delta c\|_2 = |c^\top N^{-1} c|\, \|v^* - W_0 c\|_2.$$

Taking their ratio gives

$$\frac{\|\Delta p\|_2}{\|\Delta c\|_2} = \frac{|c^\top N^{-1} p|}{|c^\top N^{-1} c|} = \frac{|\langle N^{-1/2} c, N^{-1/2} p\rangle|}{|\langle N^{-1/2} c, N^{-1/2} c\rangle|}.$$

By the theorem's assumption, there exists a constant $\lambda > 0$ such that

$$\langle N^{-1/2} c, N^{-1/2} p\rangle \geq \lambda \, \langle N^{-1/2} c, N^{-1/2} c\rangle.$$

Substituting this condition into the previous ratio yields

$$\frac{\|\Delta p\|_2}{\|\Delta c\|_2} \geq \lambda,$$

or equivalently,

$$\|\Delta W p\|_2 = \|\Delta p\|_2 \geq \lambda \|\Delta c\|_2 = \lambda \|\Delta W c\|_2.$$

□

Remark: This theorem indicates that if there exists a preserved vector p that has a large projection on the target vector c in the $N^{-1/2}$-weighted inner product space, then the perturbation on the preserved vector p may also be comparable to that on the target vector c, leading to potential performance degradation on the corresponding preserved concept.

A.3. Proof for Theorem 4.2

Proof. **Step 1: Exact preservation on the top-k subspace.** By construction $U_{2,k}^\top U_k = 0$, hence

$$\Delta W \, U_k \;=\; Z \, U_{2,k}^\top U_k \;=\; 0.$$

Therefore for any $p_\| \in \mathrm{span}(U_k)$, $W \, p_\| = (W_0 + \Delta W) \, p_\| = W_0 \, p_\|$, i.e., all top-$k$ principal directions are preserved exactly.

Step 2: The update only acts on the $(I - P_k)$-component. For a general $p \in \mathbb{R}^n$, decompose $p = p_\| + p_\perp$ with $p_\| = P_k p$ and $p_\perp = (I - P_k)p$. Since $\Delta W \, P_k = 0$,

$$(W - W_0)p \;=\; \Delta W \, p \;=\; \Delta W \, p_\perp \;=\; Z^\star U_{2,k}^\top p_\perp \;=\; Z^\star U_{2,k}^\top p.$$

Taking norms and using $\|U_{2,k}^\top p\|_2 = \|p_\perp\|_2$ gives

$$\|(W - W_0)p\|_2 \;\leq\; \|Z^\star\|_2 \, \|p_\perp\|_2.$$

Step 3: Tail bound for preserved columns. Let $p = p_i$ be a column of C_{pres}. With the thin SVD $C_{\mathrm{pres}} = U_1 \Sigma V^\top$,

$$p_i \;=\; C_{\mathrm{pres}} e_i \;=\; U_1 \Sigma V^\top e_i \;=\; \sum_{j=1}^{r} \sigma_j(C_{\mathrm{pres}}) \, v_{ij} \, u_j.$$

Since P_k projects onto $\mathrm{span}(u_1, \ldots, u_k)$, the residual is

$$\|(I - P_k)p_i\|_2^2 \;=\; \sum_{j>k} \sigma_j^2(C_{\mathrm{pres}}) \, (v_{ij})^2 \;\leq\; \sigma_{k+1}^2(C_{\mathrm{pres}}) \sum_{j>k} (v_{ij})^2 \;\leq\; \sigma_{k+1}^2(C_{\mathrm{pres}}).$$

Hence $\|(I - P_k)p_i\|_2 \leq \sigma_{k+1}(C_{\mathrm{pres}})$, and combining with Step 2 yields

$$\|(W - W_0)p_i\|_2 \;\leq\; \|Z^\star\|_2 \, \sigma_{k+1}(C_{\mathrm{pres}}),$$

as claimed in the theorem statement (with W' replaced by W in our notation). □

Remark: Our proposed method can *exactly preserve* the top-k principal directions of the preserved subspace. Moreover, for any preserved column p_i, the perturbation norm is upper bounded by the tail energy beyond the top-k singular vectors, scaled by the problem-dependent factor $\|Z^\star\|_2$. In the case where the rank of C_{pres} is less than or equal to k, the preserved set is *exactly* preserved.

A.4. Proof for Theorem 4.3

Proof. **First identity.** By the closed form (CF),

$$Z^\star C_\perp^{(k)} = B\,(C_\perp^{(k)})^\top \Big(C_\perp^{(k)}(C_\perp^{(k)})^\top\Big)^+ C_\perp^{(k)}.$$

Recall the standard pseudoinverse projection identity: for any matrix X, $X^\top(XX^\top)^+X$ is the orthogonal projector onto $\mathrm{row}(X)$. With the thin SVD $C_\perp^{(k)} = U_{q_k}^{(k)}\Sigma_{q_k}^{(k)}V_{q_k}^{(k)\top}$, this projector equals $V_{q_k}^{(k)}V_{q_k}^{(k)\top}$. Hence

$$Z^\star C_\perp^{(k)} \;=\; B\,V_{q_k}^{(k)}V_{q_k}^{(k)\top}.$$

Using $C_\perp^{(k)} = U_{2,k}^\top C_{\mathrm{tgt}}$,

$$(W - W_0)C_{\mathrm{tgt}} = \Delta W^\star C_{\mathrm{tgt}} = Z^\star U_{2,k}^\top C_{\mathrm{tgt}} = Z^\star C_\perp^{(k)} = B\,V_{q_k}^{(k)}V_{q_k}^{(k)\top},$$

which proves the first statement.

Per-column lower bound. Fix a target column index i and set $y_i := V_{q_k}^{(k)\top}e_i$. By taking the i-th column of the previous identity,

$$(W - W_0)c_i = (W - W_0)C_{\mathrm{tgt}}e_i = B\,V_{q_k}^{(k)}V_{q_k}^{(k)\top}e_i = B\,V_{q_k}^{(k)}y_i.$$

Under the hypothesis that $BV_{q_k}^{(k)}$ has full column rank, its smallest singular value $\sigma_{\min}(BV_{q_k}^{(k)})$ is strictly positive, and the standard singular-value inequality yields

$$\|(W - W_0)c_i\|_2 = \|BV_{q_k}^{(k)}y_i\|_2 \;\geq\; \sigma_{\min}(BV_{q_k}^{(k)})\,\|y_i\|_2,$$

which is exactly the claimed bound. □

Remark: Our method can exactly fit the part of the target update B that lies in the identifiable row space $\mathrm{row}(C_\perp^{(k)})$. Under a mild compatibility assumption, each target column with nonzero leverage in this row space ($\|V_{q_k}^{(k)\top}e_i\|_2 > 0$) is guaranteed to be modified by at least $\sigma_{\min}(BV_{q_k}^{(k)})$ times its leverage $\|V_{q_k}^{(k)\top}e_i\|_2$. In particular, in the single-target case ($T = 1$), if the target concept c is not contained in the top-k preserved subspace (i.e., $U_k^\top c \neq c$), then the pseudo-inverse solution gives an exact fit on the erased concept, i.e. $Wc = W_0c^\star$.

B. Detailed Experiment Settings

B.1. Object Erasing

To ensure a fair and controlled comparison across all erasure methods, we assign a fixed anchor concept to each target object category. This guarantees that UCE and DP operate under identical proxy vectors v_i^*, thereby isolating differences in performance to the erasure mechanisms themselves rather than to variations in replacement semantics. For every target concept, the chosen anchor represents a semantically neutral or structurally compatible object, enabling a clear evaluation of how effectively each method suppresses the target while redirecting the model toward the specified substitute.

Table 4. Anchor concepts used for object-level concept erasure. Each target is paired with a fixed anchor to ensure consistent proxy vectors across UCE and DP.

Target Concept	Cassette Player	Chain Saw	Church	English Springer	French Horn	Garbage Truck	Gas Pump	Golf Ball	Parachute	Tench
Anchor Concept	Box	Stick	Temple	Cat	Drum	Bus	Dispenser	Sphere	Cloth	Cucumber

The choice of anchors in Table 4 follows the suggestions by ChatGPT 4.1, by considering the semantic meanings. These anchor selections are kept consistent across all visual and quantitative evaluations. This standardized setup allows us to directly compare how different methods respond to identical replacement instructions, and it highlights the extent to which each algorithm both eliminates the target concept and preserves the integrity of non-target categories.

Note that in the above settings, we simplify the first projection step by defining the safe subspace using a single anchor vector. An exploration of more complex constructions of the safe region is provided in Appendix I.

B.2. Evaluation Protocol for Object Erasure

As alluded to earlier, our evaluation procedure is designed to be more stringent and representative of real-world semantic distinctions than those used in prior work. To reduce ambiguity in classifier predictions and diffusion outputs, we merge concept labels that are visually or semantically close. For example, categories such as "cassette player," "tape player," and "tape" are treated as a single unified concept. This consolidation avoids overcounting near-duplicate labels and results in a more accurate and conservative estimate of how well a model retains or erases a target concept. Consequently, the baseline Stable Diffusion model exhibits notably higher accuracy under our protocol, reflecting the increased ability of the unified classification task.

In addition, whereas earlier studies [8] often rely on Top-5 accuracy, we report exclusively Top-1 accuracy to provide a stricter and more discriminative measure of model behavior. This choice ensures that all reported metrics reflect precise, single-label correctness rather than broader category inclusion.

B.3. Artist Style Erasure

For each artist, we construct an extended label set to evaluate style erasure more comprehensively. Using ChatGPT-generated descriptors, each label set includes: (i) the five primary artists considered in this study, (ii) general artistic categories such as "classical art" and "modern art", and (iii) artist-specific descriptive phrases that capture characteristic stylistic elements (e.g., "dramatic lighting", "colorful landscape", "abstract portrait"). These labels allow us to assess both direct stylistic removal and potential drift toward semantically related artistic styles. The complete label sets used in our experiments are listed below.

- **Andy Warhol:** Pablo Picasso, Vincent van Gogh, Rembrandt, Andy Warhol, Caravaggio, Campbell's soup can, Marilyn Monroe portrait, screen printing, celebrity portrait, modern art, classical art.
- **Caravaggio:** Pablo Picasso, Vincent van Gogh, Rembrandt, Andy Warhol, Caravaggio, Calling of Saint Matthew, Judith Beheading Holofernes, tenebrism, dramatic lighting, religious scene, Baroque, Renaissance, 17th-century art, classical art, realism.
- **Pablo Picasso:** Pablo Picasso, Vincent van Gogh, Rembrandt, Andy Warhol, Caravaggio, Guernica, Blue Period, Rose Period, African mask, guitar collage, abstract portrait, geometric art, modern art, classical art.
- **Rembrandt:** Pablo Picasso, Vincent van Gogh, Rembrandt, Andy Warhol, Caravaggio, The Night Watch, self-portrait, Saskia portrait, chiaroscuro, Dutch master, Baroque, classical art, impressionism, cubism, modern art.
- **Vincent van Gogh:** Pablo Picasso, Vincent van Gogh, Rembrandt, Andy Warhol, Caravaggio, generic impressionist painting, abstract expressionism, post-impressionist art, colorful landscape, Starry Night scene, sunflower painting, wheat field artwork, cypress trees, countryside scene, generic modern art, unspecified artist style.

C. Imperfection of Preservation

C.1. The Impacts of Positional Embedding

Although the DP algorithm theoretically enforces orthogonality between erased and preserved subspaces, *perfect preservation* of non-target concepts is not always achieved in practice. This discrepancy primarily arises from the *positional embedding structure* in diffusion models, where each token embedding is not used in isolation but is *summed* with its positional encoding before entering the attention and MLP layers.

Formally, let the raw content embedding for a token be $c_i \in \mathbb{R}^n$ and its positional embedding be $q_i \in \mathbb{R}^n$. The effective input to the model is then

$$z_i = e_i + q_i. \tag{16}$$

During concept erasure, DP computes an update ΔW satisfying the preservation constraint

$$\Delta W e_i = 0, \tag{17}$$

which guarantees that all preserved content embeddings C_{pres} remain unaffected in the ideal case. However, in the actual model, the transformation is applied to the fused embedding z_i, not to c_i alone.

Since q_i is not fixed (the word can appear at arbitrary location) and generally *not orthogonal* to the erased directions, the effective transformation satisfies

$$\Delta W z_i \neq 0. \tag{18}$$

This residual term introduces a small coupling between erased and preserved subspaces, leading to the minor performance drop observed empirically. Note the non-target concepts can appear in any position, and in general, it is not feasible to also require $\Delta W q_i = 0$ for all q_i.

Importantly, this limitation is *not unique to DP*. Closed-form projection methods such as UCE are subject to the same positional interaction, since they also operate in the linearized embedding space and do not explicitly disentangle positional components. In other words, while both DP and UCE guarantee subspace orthogonality for pure content embeddings, the *additive nature of positional encodings* inherently prevents perfect preservation in diffusion architectures.

C.2. Self-Attention in Encoder

Specifically, the CLIP text encoder used in diffusion models applies multiple *self-attention blocks* when producing text embeddings, so the resulting embedding of each token is no longer independent of the others. As a result, token representations become contextualized and partially mixed across the prompt. For example, in the prompt "An image of Church", the embedding associated with "Church" after encoding is not merely the isolated concept embedding of "Church", but a contextualized representation that also carries weak information from the surrounding tokens through self-attention. Consequently, even if ΔW is constructed to be orthogonal to non-target concept embeddings in principle, the actual encoded representations processed by the model may still be slightly perturbed. Together with the effect of positional embeddings, this token mixing provides a practical explanation for the small but consistent deviations from perfect preservation observed in our experiments.

C.3. Why Editing on Embedding Layers Produces Better Preservation.

Operating directly on the *embedding layer* of the encoder avoids the positional–intervene concepts before the positional embedding and self-attention. At the embedding layer, the model processes the content vectors c_i *before* they are fused with positional embeddings. This allows the preservation constraint to be enforced exactly.

When concept erasure is applied at the embedding layer, the update ΔW_{emb} acts only on c_i:

$$z_i = (W_{\text{emb}} + \Delta W_{\text{emb}})\, c_i + q_i.$$

The preservation condition becomes

$$\Delta W_{\text{emb}}\, C_{\text{pres}} = 0,$$

which directly implies

$$(W_{\text{emb}} + \Delta W_{\text{emb}})\, C_{\text{pres}} = W_{\text{emb}}\, C_{\text{pres}}.$$

Since positional embeddings are added *after* the content projection, they do not interfere with this constraint. The effective representation remains

$$z_i' = W_{\text{emb}}\, C_{\text{pres}} + q_i,$$

which is identical to the original representation for all preserved concepts.

These results explain why embedding-level editing consistently yields more stable preservation behavior: it achieves exact orthogonality for content embeddings, results in cleaner and more localized updates, and eliminates interference caused by positional encodings, as further demonstrated in our FLUX visualizations (Appendix H).

D. Why UCE Underperforms on *Cassette Player* and *Golf Ball*

D.1. Last Word Erasing

Although UCE generally provides strong erasure performance, we observe two notable failure cases in our experiments: *Cassette Player* and *Golf Ball*. Upon closer inspection, these failures arise from the way UCE constructs the concept embedding used for editing.

UCE uses only the last token embedding. In the official implementation of UCE, the concept embedding for a multi-word prompt is constructed by selecting *only the last token* of the prompt. The relevant code snippet from the official release is shown below:

```
t_emb = pipe.encode_prompt(
        prompt=e,
        device=device,
        num_images_per_prompt=1,
        do_classifier_free_guidance=False)

last_token_idx = (
        pipe.tokenizer(
                e,
                padding="max_length",
                max_length=pipe.tokenizer.model_max_length,
                truncation=True,
                return_tensors="pt",
        )["attention_mask"]
).sum() - 2

uce_erase_embeds[e] = t_emb[0][:, last_token_idx, :]
```

In particular, the above codes effectively select the "last token index" for all target concept. For many artistic concepts such as "Van Gogh" or "Picasso", this design choice is relatively benign because the semantic meaning is concentrated in the final token. However, for compound nouns commonly found in the object-erasing benchmark, the last token does *not* capture the dominant semantics.

Why this fails for "Cassette Player" and "Golf Ball". In both of these categories, the first token carries the primary semantic load: "cassette" in "cassette player" and "golf" in "golf ball". UCE, however, replaces only the second token. For example:

- Replacing `ball` with `sphere` leads to prompts interpreted by the model as "golf sphere", which often continues to produce golf-ball–like objects. Such outputs remain highly classifiable as *golf ball* by the pretrained ResNet-50 classifier.
- Similarly, replacing `player` in `cassette player` fails to remove the defining visual features associated with the first token, causing the resulting images to retain the appearance of a cassette-like object.

This explains the substantially higher erasing accuracy for UCE on these two categories reported in Table 1.

Replacing all tokens improves UCE in these cases. For completeness, we run an additional experiment in which UCE replaces the embeddings of *all* tokens in the target phrase rather than only the last one. Under this corrected setting:

- The erased accuracy for *Golf Ball* improves dramatically, decreasing from 12.0 to 2.0, which is comparable to our DP method.
- The accuracy drop on non-target concepts is also reduced, improving from 20.8 to 16.2.

D.2. DP Still Achieves Better Preservation

Despite these improvements, UCE still introduces substantially larger perturbations to non-target concepts. Under the same corrected setting, DP achieves a much lower preservation drop of only 3.3, demonstrating that even with improved token handling, UCE's single-projection update remains more disruptive to unrelated concept directions.

This analysis confirms that UCE's underperformance is primarily due to its reliance on the last-token embedding, and that our DP method not only avoids this limitation but also maintains significantly better preservation of non-target concepts.

E. Other Metrics for Image Assessment

Beyond classification-based accuracy metrics used in the main experiments, we further evaluate the visual quality and perceptual fidelity of generated images of the FLUX model using several widely adopted generative-model metrics: LPIPS [71], PSNR [6], SSIM [62], and FID [31].

LPIPS measures perceptual similarity using deep feature distances, providing sensitivity to semantic changes in image content. PSNR and SSIM quantify pixel-level and structural similarity, respectively, enabling assessment of how closely the edited outputs preserve low-level visual attributes. FID evaluates realism at the distribution level by comparing feature statistics of generated images to those of real images. Together, these metrics offer a complementary perspective on the impact of concept erasure, allowing us to assess not only whether the target concept is successfully suppressed, but also how strongly each method affects the overall perceptual quality and statistical properties of non-target generations.

It is important to emphasize that these metrics are evaluated **only on the non-target concepts**. Measures such as LPIPS [71], PSNR [6], SSIM [62], and FID [31] quantify differences between images generated *before and after* concept erasure, and therefore assume that the underlying semantic content should remain consistent across the two states. This assumption naturally holds for non-target concepts, where the objective is to preserve visual fidelity and minimize unintended perturbations.

In contrast, applying these metrics to the *target* concepts would be inappropriate, since concept erasure is explicitly designed to alter (and ideally remove) the original content. The images before and after erasure are thus expected to differ substantially, rendering such reconstruction-based metrics neither meaningful nor interpretable for evaluating erasure quality.

Table 5. Comparison of DP and UCE across preserved concepts using **LPIPS**, **PSNR**, **SSIM**, and **FID**. Lower is better for LPIPS and FID; higher is better for PSNR and SSIM.

Concept	**LPIPS** ↓		**PSNR** (dB) ↑		**SSIM** ↑		**FID** ↓	
	UCE	DP	UCE	DP	UCE	DP	UCE	DP
Cassette Player	$0.1170_{\pm 0.0287}$	$\mathbf{0.0506}_{\pm 0.0248}$	$19.60_{\pm 2.47}$	$\mathbf{24.61}_{\pm 4.45}$	$0.8019_{\pm 0.0580}$	$\mathbf{0.8883}_{\pm 0.0559}$	14.29	**8.57**
Chain Saw	$0.1113_{\pm 0.0262}$	$\mathbf{0.0362}_{\pm 0.0339}$	$19.79_{\pm 2.47}$	$\mathbf{26.63}_{\pm 4.90}$	$0.8060_{\pm 0.0600}$	$\mathbf{0.9141}_{\pm 0.0518}$	11.87	**5.44**
Church	$0.1222_{\pm 0.0282}$	$\mathbf{0.0388}_{\pm 0.0208}$	$19.33_{\pm 2.41}$	$\mathbf{25.73}_{\pm 4.24}$	$0.7964_{\pm 0.0617}$	$\mathbf{0.9091}_{\pm 0.0461}$	14.25	**6.86**
English Springer	$0.1225_{\pm 0.0397}$	$\mathbf{0.0836}_{\pm 0.0378}$	$19.36_{\pm 2.51}$	$\mathbf{21.68}_{\pm 3.19}$	$0.7911_{\pm 0.0608}$	$\mathbf{0.8404}_{\pm 0.0600}$	14.74	**11.28**
French Horn	$0.1211_{\pm 0.0317}$	$\mathbf{0.0367}_{\pm 0.0202}$	$19.75_{\pm 2.40}$	$\mathbf{26.43}_{\pm 4.03}$	$0.7995_{\pm 0.0646}$	$\mathbf{0.9110}_{\pm 0.0483}$	14.92	**6.76**
Garbage Truck	$0.1204_{\pm 0.0396}$	$\mathbf{0.0465}_{\pm 0.0313}$	$19.81_{\pm 2.50}$	$\mathbf{25.29}_{\pm 3.73}$	$0.8040_{\pm 0.0668}$	$\mathbf{0.8998}_{\pm 0.0616}$	14.56	**7.99**
Gas Pump	$0.1155_{\pm 0.0301}$	$\mathbf{0.0470}_{\pm 0.0317}$	$19.69_{\pm 2.36}$	$\mathbf{25.29}_{\pm 4.67}$	$0.8009_{\pm 0.0610}$	$\mathbf{0.8951}_{\pm 0.0599}$	13.07	**7.59**
Golf Ball	$0.1228_{\pm 0.0268}$	$\mathbf{0.0362}_{\pm 0.0123}$	$18.98_{\pm 1.63}$	$\mathbf{25.16}_{\pm 1.96}$	$0.7885_{\pm 0.0498}$	$\mathbf{0.9098}_{\pm 0.0254}$	15.30	**6.90**
Parachute	$0.1159_{\pm 0.0337}$	$\mathbf{0.0682}_{\pm 0.0192}$	$19.45_{\pm 2.50}$	$\mathbf{22.34}_{\pm 2.56}$	$0.7958_{\pm 0.0599}$	$\mathbf{0.8603}_{\pm 0.0395}$	14.54	**9.93**
Tench	$\mathbf{0.1121}_{\pm 0.0287}$	$0.1345_{\pm 0.0588}$	$\mathbf{19.66}_{\pm 2.48}$	$18.99_{\pm 3.24}$	$\mathbf{0.8016}_{\pm 0.0633}$	$0.7769_{\pm 0.0869}$	**13.61**	14.97
Average	**0.1181**	0.0578	19.54	**24.22**	0.7996	**0.8805**	14.08	**8.73**

Table 5 reports a comprehensive comparison between DP and UCE across ten preserved concepts, evaluated using LPIPS, PSNR, SSIM, and FID. For metrics where lower values indicate better performance (LPIPS and FID), DP consistently outperforms UCE on nine out of ten concepts. The only exception is the "Tench" class, where UCE achieves a slightly lower LPIPS score. On average, DP achieves a substantially lower LPIPS score (0.0578 vs. 0.1181), indicating a significantly improved perceptual similarity to the target images.

For distortion-based metrics where higher values indicate better image fidelity (PSNR and SSIM), DP again demonstrates favorable behavior. DP achieves higher PSNR and SSIM values on all concepts except "Tench", showing a robust improvement in reconstruction fidelity. Averaged across all concepts, DP improves PSNR by approximately $+4.7$ dB over UCE (24.22 vs. 19.54) and achieves a higher SSIM score (0.8805 vs. 0.7996), demonstrating consistently better structural alignment and visual coherence.

In terms of generative quality, DP achieves notably lower FID scores on nine out of ten concepts, again with the sole exception of "Tench". The average FID of DP (8.73) is substantially lower than that of UCE (14.08), indicating that DP produces more realistic and distribution-consistent outputs.

Overall, the results show that *DP outperforms UCE across all four metrics and on nearly all individual concepts*. This demonstrates that DP provides superior perceptual similarity, lower distortion, higher structural fidelity, and more realistic generative quality when preserving concept-specific image content.

F. Additional Experiments on SD 1.5

While the main paper focuses on Stable Diffusion 1.4 due to its widespread use in prior concept-erasure research and its role as a canonical benchmark, we also conduct a parallel set of experiments on Stable Diffusion 1.5 to assess the robustness and generality of our approach. The SD 1.5 backbone differs from SD 1.4 in both training distribution and visual appearance characteristics, making it a meaningful testbed for evaluating consistency across model variants.

Table 6. Results on SD 1.5 for all algorithms. Each block includes both original and post-update accuracies. Left: **Target Class** shows erasure performance (**Erased Accuracy** ↓). Right: **Other Classes** reports the accuracy of preserved concepts (**Preservation Drop** ↓). Lower values indicate stronger erasure and better preservation.

Object	**Target Class**	**Erased Accuracy (%)** ↓					**Other Classes**	**Preservation Drop (%)** ↓				
	Original	ESD	CP	AGE	UCE	DP	Original	ESD	CP	AGE	UCE	DP
Cassette Player	60.0	$12.0_{\pm 1.0}$	$4.0_{\pm 0.5}$	$33.0_{\pm 3.7}$	41.0	**0.0**	88.8	$25.9_{\pm 2.1}$	$27.4_{\pm 1.9}$	$6.3_{\pm 0.4}$	20.4	**4.2**
Chain Saw	76.0	$1.0_{\pm 0.0}$	$\mathbf{0.0}_{\pm 0.0}$	$2.0_{\pm 0.5}$	**0.0**	**0.0**	87.0	$18.1_{\pm 1.8}$	$29.1_{\pm 2.2}$	$1.6_{\pm 0.2}$	2.3	**0.2**
English Springer	95.0	$1.0_{\pm 0.0}$	$\mathbf{0.0}_{\pm 0.0}$	$\mathbf{0.0}_{\pm 0.0}$	**0.0**	**0.0**	84.9	$43.9_{\pm 2.4}$	$32.6_{\pm 1.9}$	$1.9_{\pm 0.3}$	2.1	**−0.2**
Parachute	93.0	$5.0_{\pm 0.7}$	$5.0_{\pm 0.5}$	$2.0_{\pm 0.5}$	**0.0**	**0.0**	85.1	$10.4_{\pm 1.1}$	$37.7_{\pm 2.8}$	$1.3_{\pm 0.2}$	**0.3**	**0.3**
French Horn	99.0	$2.0_{\pm 0.3}$	$\mathbf{0.0}_{\pm 0.0}$	$3.0_{\pm 1.0}$	**0.0**	1.0	84.4	$17.1_{\pm 1.6}$	$29.3_{\pm 2.3}$	$4.7_{\pm 0.4}$	**1.6**	2.4
Golf Ball	100.0	$17.0_{\pm 1.5}$	$12.0_{\pm 1.0}$	$2.0_{\pm 0.3}$	61.0	**0.0**	84.3	$16.2_{\pm 1.3}$	$27.0_{\pm 2.0}$	$4.9_{\pm 0.5}$	7.6	**4.8**
Garbage Truck	93.0	$1.0_{\pm 0.3}$	$\mathbf{0.0}_{\pm 0.0}$	$11.0_{\pm 1.7}$	**0.0**	**0.0**	85.1	$18.0_{\pm 1.5}$	$39.2_{\pm 3.1}$	$1.1_{\pm 0.2}$	−1.7	**−1.9**
Tench	80.0	$\mathbf{0.0}_{\pm 0.0}$	$\mathbf{0.0}_{\pm 0.0}$	$8.0_{\pm 1.7}$	**0.0**	**0.0**	86.6	$23.6_{\pm 2.2}$	$25.4_{\pm 1.7}$	$2.9_{\pm 0.3}$	6.0	**2.2**
Gas Pump	76.0	$10.0_{\pm 1.0}$	$2.0_{\pm 1.0}$	$\mathbf{1.3}_{\pm 0.3}$	3.0	4.0	87.0	$13.6_{\pm 1.3}$	$34.8_{\pm 2.8}$	$1.6_{\pm 0.2}$	4.6	**1.4**
Church	88.0	$25.0_{\pm 2.0}$	$\mathbf{0.0}_{\pm 0.0}$	$13.0_{\pm 1.0}$	**0.0**	2.0	85.8	$27.0_{\pm 2.4}$	$35.8_{\pm 3.3}$	$4.6_{\pm 0.4}$	8.1	**4.3**
Mean	86.0	7.4	2.3	7.5	10.5	**0.7**	85.9	21.4	31.8	3.1	5.1	**1.8**

The results on Stable Diffusion 1.5 in Table 6 exhibit trends consistent with those observed for SD 1.4, further confirming that DP generalizes effectively across different diffusion backbones. Across all ten evaluated object categories, DP achieves the *lowest mean erased accuracy (0.7%)*, outperforming all competing baselines by a substantial margin. In many cases, including *Cassette Player*, *Chain Saw*, *English Springer*, *Parachute*, *Golf Ball*, *Garbage Truck*, and *Tench*, DP completely suppresses the target object, achieving a residual accuracy of 0.0%. Even in more challenging categories such as *Church* and *Gas Pump*, DP remains competitive, demonstrating that the double-projection mechanism continues to yield effective erasure despite architectural differences between SD 1.4 and SD 1.5.

In terms of preserving non-target concepts, DP again provides the strongest performance. Iterative or pruning-based approaches such as ESD and CP introduce substantial collateral degradation, often exceeding a large preservation drop. UCE performs better but still yields an average drop of 5.1%. In contrast, DP maintains an average degradation of only 1.8%, several times lower than any other method. In multiple categories, including *English Springer*, *Garbage Truck*, and *Tench*, DP results in slightly *negative* preservation drop, indicating that the overall accuracy for other non-target concepts increases.

Overall, two clear behavioral clusters emerge. Methods like CP and ESD display high variance and significant unintended perturbations due to their reliance on broad, iterative parameter modifications. UCE performs reasonably on simpler single-token concepts but struggles with multi-token cases (e.g., *Golf Ball*, *Cassette Player*), reflecting the token-selection limitations discussed previously. By contrast, DP remains uniformly stable: its closed-form update isolates the erasure direction while explicitly preserving the orthogonal subspace, enabling it to maintain high fidelity even when concept representations are semantically entangled.

These findings reinforce the central message of this work: DP provides strong, architecture-agnostic concept erasure while consistently minimizing unintended degradation, validating the robustness of the proposed double-projection framework across both classical and updated diffusion model variants.

G. Visualization for Stable Diffusion 1.4

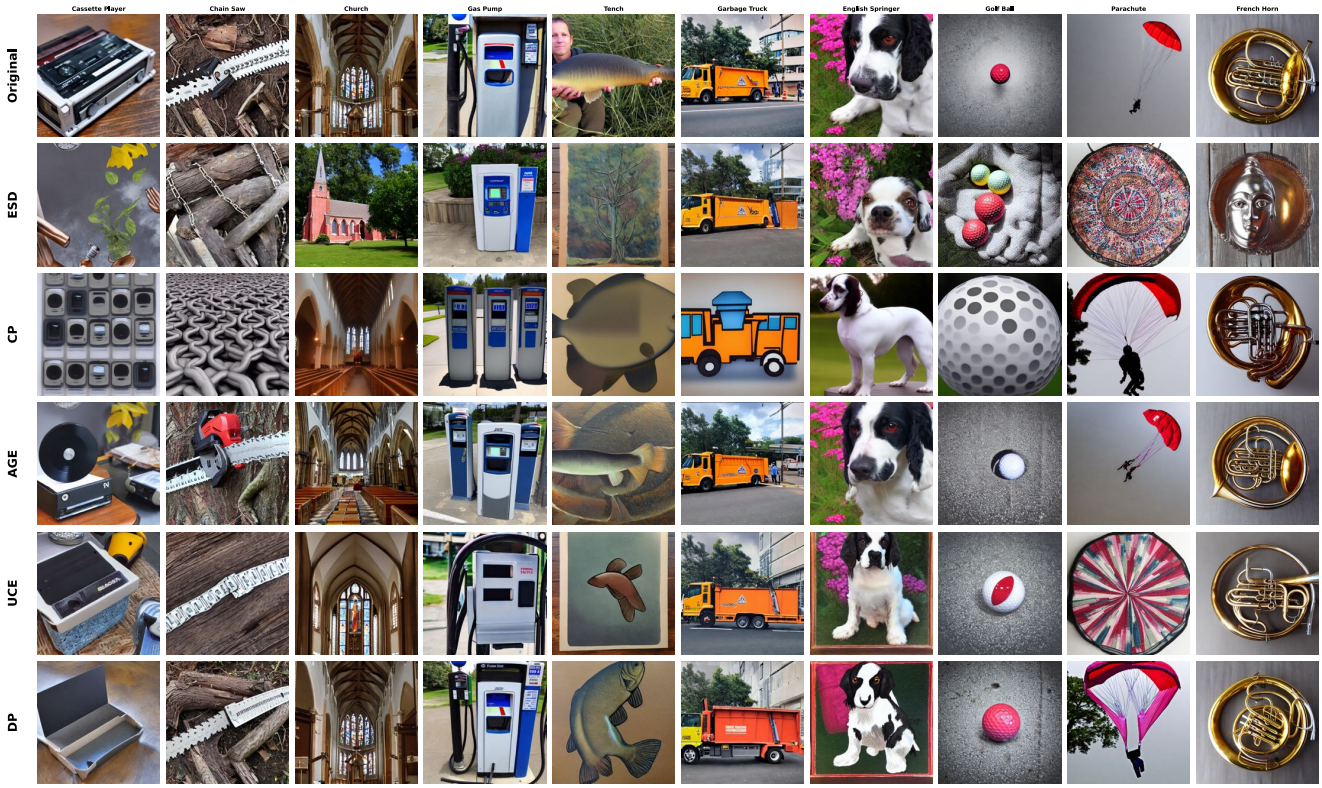


Figure 3. Concept erasure on "Cassette Player" with anchor concept "Box". The first column shows the target concept to be erased.

Figure 3 presents qualitative visualizations for all five erasure methods, ESD, CP, AGE, UCE, and the proposed DP, using the cassette player category as the target concept. For each method, we display the first generated sample from the ten evaluated categories, with the objective of suppressing the target concept in the first column while leaving the remaining nine concepts unaffected.

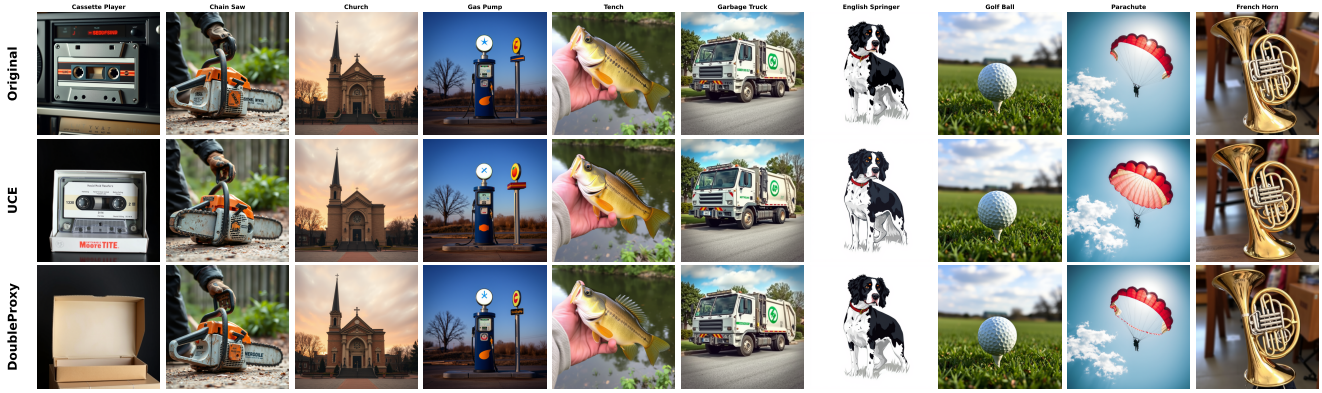
For the target concept, the closed-form approaches UCE and DP both succeed in preserving the overall structural layout of the original image while substituting the target semantics with the designated anchor concept. Notably, DP produces outputs that align more faithfully with the anchor concept box, yielding clearer and more coherent substitutions than those produced by UCE. This behavior visually corroborates the quantitative results reported earlier, where DP demonstrated stronger erasure performance on the target class.

For the nine non-target concepts, however, all methods exhibit some degree of perturbation. These deviations are especially pronounced for CP, whose outputs diverge substantially from the original images, indicating weaker preservation capability. DP also shows mild perturbations on non-target categories, though the changes are considerably smaller and do not alter the primary semantics of the generated content.

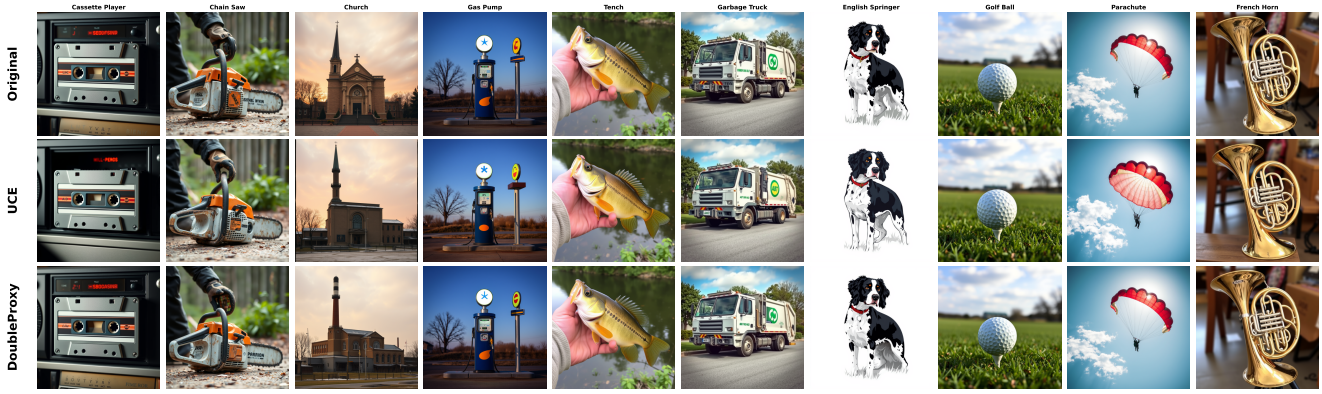
In contrast, when concept erasure is applied directly to the embedding layer, as demonstrated in FLUX (see Appendix H), the model preserves non-target concepts much more reliably. This comparison suggests that interventions performed within deeper architectural components, such as attention blocks, are more likely to propagate unintended changes throughout the network. Even with closed-form constraints, edits at these deeper layers can influence representations beyond the targeted concept.

H. Visualization for FLUX

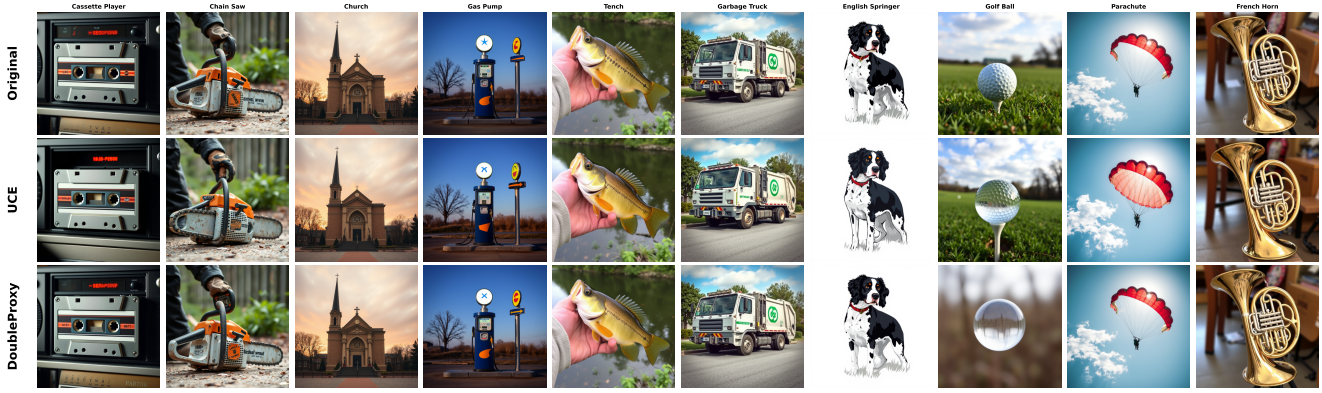
To further illustrate the qualitative behavior of concept erasure, Figure 4 presents visualizations for three representative target concepts—cassette player, church, and golf ball. For fairness and controlled comparison, each target concept is paired with a fixed anchor concept used as the replacement proxy v_i^*. Across all examples, the proposed DP method consistently removes



(a) Concept erasure on “Cassette Player” with anchor concept “Box”. The first column shows the target concept to be erased.



(b) Concept erasure on “Church” with anchor concept ”Factory”. The third column shows the target concept to be erased.



(c) Concept erasure on “Golf Ball” with anchor concept ”Sphere”. The eighth column shows the target concept to be erased.

Figure 4. Concept erasure on a few target concepts with FLUX. Results demonstrate that the proposed DP method successfully suppresses the target concept while generating images faithful to the replaced anchor concept.

the target concept while producing images that align closely with the intended anchor semantics. In contrast, UCE often retains recognizable traces of the original concept, indicating incomplete suppression.

This difference is most evident in Figure 4b: although both methods attempt to erase the concept church using factory as the anchor, DP produces structures that clearly resemble industrial buildings, whereas UCE-generated images continue to exhibit architectural features characteristic of churches. Similar patterns appear across the remaining examples—DP reliably redirects the model’s output toward the anchor concept, while UCE frequently preserves residual cues associated with the target. These qualitative results reinforce our quantitative findings, demonstrating that DP achieves more effective concept removal and cleaner semantic substitution, thereby validating its superior erasure capability.

I. Ablation Studies

I.1. Ablation on the First Projection

The first component of the DP framework is the first projection step, where the target concept is mapped into a user-defined safe subspace. In the main paper, we adopt a simplified configuration in which the safe subspace is defined by a single anchor concept, mirroring the setup used in UCE to ensure a fair comparison. However, the DP formulation naturally supports larger and more expressive safe regions, which may improve preservation fidelity or alter erasure behavior depending on the geometry of the selected subspace.

To illustrate this effect, we perform an ablation study on the target concept "Church". Specifically, we compare two settings:

1. A multi-vector safe region constructed from the concepts "tower" and "factory".
2. A single-vector safe region using only "factory" as the anchor (as in the main experiments).

Table 7. Ablation study on the construction of the safe subspace for the target concept *Church*. We compare a multi-vector safe region (Tower + Factory) with a single-vector anchor (Factory). Left block reports erasure performance; right block reports preservation quality on non-target classes.

Method	**Target Concept: Church**			**Other Classes (Avg.)**		
	Original	**After Erasure ↓**	**Drop ↑**	**Original**	**After Erasure ↑**	**Drop ↓**
Factory only	99.0%	12.0%	87.0%	89.7%	89.4%	0.3%
Tower + Factory	99.0%	3.0%	96.0%	89.7%	89.6%	0.1%

Table 7 demonstrates that broader safe subspaces yield more effective erasure while preserving non-target concepts with minimal degradation. Even so, using a single anchor vector often remains the preferred strategy in practice due to its simplicity and ease of deployment.

I.2. Ablation on the Second Projection

The role of the second projection can be directly assessed by comparing our method with UCE (e.g, Tables 1). Since both approaches use the *same* anchor vectors v_i^* and differ only in the presence of the nullspace projection, these results naturally serve as ablation studies isolating the contribution of the second projection. This performance gap highlights the necessity of the second projection: without restricting updates to the left nullspace of preserved embeddings, as in UCE, concept removal introduces noticeable interference to unrelated representations. In contrast, enforcing the nullspace constraint ensures that modifications remain geometrically orthogonal to the preserved subspace, resulting in significantly more stable and predictable behavior across both diffusion and flow-matching architectures.

J. Generation on Other Classes

To more comprehensively evaluate the generality of our concept-erasure framework, we conduct an additional set of experiments on a broader collection of ImageNet classes beyond these ten categories used in the main paper. In particular, we focus on the FLUX model for UCE and DP methods. These experiments serve two primary purposes. First, they allow us to examine the stability of our method when applied across a wider range of visual concepts with diverse semantics and visual structures. Second, they enable a deeper analysis of how preservation quality behaves when the preserved concept matrix C_{pres} contains classes that differ in similarity to the target concept.

To construct this extended benchmark, we curated a set of seven ImageNet-confirmed synsets spanning multiple semantic domains, including household objects, animals, vehicles, furniture, and sports equipment. The selected classes are: `coffee_mug`, `beer_bottle`, `African_elephant`, `airliner`, `mountain_bike`, `loudspeaker`, and `volleyball`. This selection follows the suggestions from ChatGPT and ensures broad coverage across the ImageNet hierarchy while avoiding redundancy among preserved concepts.

Notably, we intentionally include `loudspeaker`, which is semantically related to "Cassette Player". By doing so, we create a more challenging scenario for evaluating the behavior of C_{pres}: the preservation matrix now contains a near-neighbor of the erased concept, allowing us to test whether the erasure update can suppress only the target direction without unintentionally diminishing representations associated with semantically adjacent classes. The remaining concepts, chosen to be visually and semantically distinct from the target, provide a stable set for assessing preservation fidelity.

Class	**Original**	**UCE**		**DP**	
		Preserved ↑	Drop ↑	Preserved ↑	Drop ↑
African elephant	85.0%	79.0%	+6.0%	84.0%	**+1.0%**
airliner	96.0%	94.0%	+2.0%	96.0%	**+0.0%**
beer bottle	96.0%	91.0%	+5.0%	96.0%	**+0.0%**
coffee mug	71.0%	66.0%	+5.0%	70.0%	**+1.0%**
loudspeaker	93.0%	95.0%	-2.0%	96.0%	**-3.0%**
mountain bike	100.0%	98.0%	+2.0%	100.0%	**+0.0%**
volleyball	39.0%	26.0%	+13.0%	35.0%	**+4.0%**
Mean	82.86%	78.43%	4.43%	82.43%	**0.43%**

Table 8. Classification accuracy comparison on general ImageNet classes before and after concept erasure on "Cassette Player". "Original" denotes accuracy on the unmodified model. "Preserved" is accuracy after applying UCE or DP. "Drop" is defined as (Original − Preserved), where smaller drops (bold) indicate better preservation of general concepts.

To assess whether concept erasure affects recognition performance on unrelated classes, we evaluate the classification accuracy on seven general ImageNet categories (Table 8). Since the original model predictions are identical for both methods, we report them only once and compare the post-erasure accuracy ("Preserved") as well as the accuracy drop (Original − Preserved). A smaller drop indicates better retention of general concepts unrelated to the targeted erased concepts.

Across the seven categories, DP consistently exhibits smaller drops in accuracy, achieving an average drop of only 0.43%, compared to 4.43% for UCE. DP matches or outperforms UCE on every class, including "loudspeaker" where the drop is negative, indicating an unexpected boost in accuracy after applying the method. In contrast, UCE frequently induces substantial degradation, most notably on the "volleyball" class where the accuracy falls by 13 percentage points.

The preserved accuracies further support this trend: DP retains an average of 82.43% classification accuracy post-erasure, nearly identical to the original value of 82.86%. UCE, however, drops to an average of 78.43%, showing that the method introduces notable unintended interference in general recognition capabilities.

Since the target concept in our experiments is "cassette player" (the target concept is chosen alphabetically.), it is natural to examine how erasure interacts with semantically related categories. Among the evaluated classes, "loudspeaker" is arguably the closest in terms of object type and visual context: both involve audio equipment, share similar geometric structures, and frequently co-occur in similar environments. One might reasonably expect such conceptual proximity to induce a noticeable decline in recognition performance after erasure.

However, the empirical results reveal that the influence on "loudspeaker" is remarkably minor for both methods. DP exhibits only a -3.0% drop, while UCE shows a slightly smaller -2.0% drop. Importantly, both drops are negative, indicating that recognition accuracy actually *improves* after concept removal. This suggests that the removed "cassette player" features are sufficiently specialized and do not interfere with the broader representation needed to recognize a "loudspeaker". The fact that DP maintains robust performance on this semantically adjacent class, while still achieving the intended erasure, highlights its ability to localize the targeted concept without degrading conceptually overlapping regions of the feature space.

Overall, these results demonstrate that DP generalizes more safely: it removes the targeted concept while preserving recognition performance on unrelated classes, whereas UCE exhibits measurable collateral damage across diverse ImageNet categories.