Stake the Points: Structure-Faithful Instance Unlearning

Supplementary Material

A. Baselines

We compared our method, STRUCTGUARD, with representative unlearning approaches [6, 12, 37] aligned with the misclassification objective and the instance-level unlearning scenario:

- **FISHER** [12]: Estimates the importance of parameters for the forget set and suppresses influential weights according to their effect.
- **NEGGRAD** [12]: Applies gradient reversal on the forget data to directly drive the model away from the forgotten information.
- **RAWP** (a variant of AWP [37]): Repeatedly perturbs model weights using forget data to induce misclassification by destabilizing weights linked to forgotten content.
- **L2UL** [6]: Uses adversarial variants of the forget samples to prevent representation-level forgetting, and constrain parameters sensitive to the forget samples through parameter-level forgetting.
- **ADV** (a variant of L2UL [6]): Performs gradient ascent on the forget set and incorporates adversarial examples generated from it to maintain the retained representations.

B. Class-level Unlearning

Although our focus is on instance-level unlearning scenarios, we also evaluated whether the proposed method extends to class-level deletion. Following prior studies [22], we designated 10% of the entire set of classes as the forget set and conducted experiments on CIFAR-10 and CIFAR-100. The results are summarized in Table 7. All methods successfully remove the designated classes on both datasets, except for FISHER on CIFAR-10. In particular, FISHER and NEGGRAD exhibit unstable performance on both datasets, because class-level deletion requires removing all samples belonging to the designated classes, which intensifies structural collapse and hinders retention. RAWP maintains performance on CIFAR-10 but struggles on CIFAR-100 as the number of forgotten classes grows. ADV demonstrates stronger retention than RAWP on CIFAR-100, due to its use of adversarial samples for retention. L2UL consistently outperforms all baselines on both datasets, making it the strongest baseline for class-level deletion. Notably, compared to L2UL, our approach achieves average improvements of 10.45% on the test-set accuracy $\mathcal{A}_{\text{test}}$ and 12.28% on the retention-set accuracy $\mathcal{A}_r$, averaged across both datasets. These noticeable performance gaps reflect the greater structural collapse caused by representation distortion under class-level deletion, underscoring that preserving the relational structure of the retained knowledge becomes even more critical.

Table 7. Results of the class-level unlearning scenario.

	Method	CIFAR-10	CIFAR-100
$\mathcal{A}_{\text{test}}(\uparrow)$	BEFORE	92.59	77.10
	ORACLE	79.14	57.44
	FISHER	11.84	1.92
	NEGGRAD	13.77	20.82
	RAWP	59.31	28.14
	ADV	62.20	42.53
	L2UL	64.56	43.38
	STRUCTGUARD	**78.20**	**50.65**
$\mathcal{A}_r(\uparrow)$	BEFORE	99.63	99.98
	ORACLE	93.06	87.66
	FISHER	12.62	1.03
	NEGGRAD	14.94	26.60
	RAWP	69.14	39.98
	ADV	74.22	63.38
	L2UL	75.79	64.32
	STRUCTGUARD	**93.48**	**71.19**
$\mathcal{A}_f(\uparrow)$	BEFORE	0.00	0.00
	ORACLE	100.00	100.00
	FISHER	93.68	100.00
	NEGGRAD	100.00	100.00
	RAWP	100.00	100.00
	ADV	100.00	100.00
	L2UL	100.00	100.00
	STRUCTGUARD	100.00	100.00

Table 8. Comparison of different semantic encoders under $k = 256$.

Encoder	CIFAR-10			CIFAR-100		
	$\mathcal{A}_{\text{test}}(\uparrow)$	$\mathcal{A}_r(\uparrow)$	$\mathcal{A}_f(\uparrow)$	$\mathcal{A}_{\text{test}}(\uparrow)$	$\mathcal{A}_r(\uparrow)$	$\mathcal{A}_f(\uparrow)$
SBERT	47.22	52.64	100.00	56.66	82.63	100.00
SigLIP	53.99	59.03	100.00	56.90	82.75	100.00
CLIP (ours)	**56.32**	**61.67**	100.00	**56.91**	**83.30**	100.00

C. Semantic Encoder

We conducted an analysis to examine how the choice of a semantic encoder influences unlearning. We selected two representative alternatives to CLIP: Sentence-BERT (SBERT) [28] and SigLIP [41]. SBERT is a language-only model that generates sentence embeddings, allowing us to test whether anchors derived solely from linguistic semantics can still guide structure preservation. SigLIP is a multimodal encoder trained with a sigmoid-based contrastive objective, serving as a counterpart to CLIP.

As shown in Table 8, SBERT shows the weakest performance among the alternatives, as it lacks alignment with

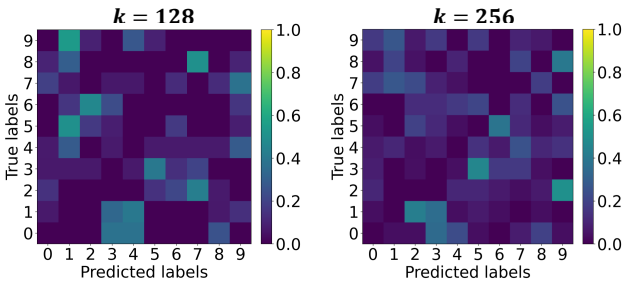

Figure 8. Confusion matrices on CIFAR-10 under $k = 128$ and $k = 256$. Each matrix shows post-unlearning predictions on D_f, with color brightness proportional to prediction frequency for each label pair.

visual information. Furthermore, SigLIP achieves higher performance than SBERT on both datasets and reaches performance close to CLIP on CIFAR-100. Because SigLIP is trained on paired image–text data, its anchors are more closely aligned with the visual representations, which facilitates structure preservation, especially when the classes exhibit rich and diverse semantics. Among the semantic encoders, CLIP provides the strongest performance within our framework. This is because the global contrastive training objective used in CLIP produces representations that align more naturally with the anchor–instance relational structure. Notably, across all semantic encoders, ours maintains a favorable deletion–retention trade-off and generalizes well, consistently outperforming all the alternatives; the baseline results are reported in Table 2 in the main paper.

D. Misclassification

To test whether our method induces the Streisand effect [12], referring to the unintended exposure of information that should have been forgotten, we examined whether forgotten samples exhibit specific prediction tendencies, as shown in Figure 8. We visualized the prediction tendencies of the unlearned model on the forget set D_f using confusion matrices at $k = 128$ and $k = 256$ to capture how its outputs deviate from the true labels. In both confusion matrices, no consistent pattern emerges; the predictions do not form any meaningful arrangement, and this becomes even less organized at $k = 256$, indicating no usable regularity that could expose forgotten information. These results show that our method prevents information leakage of forgotten instances and avoids predictable misclassification, thereby strengthening privacy.

E. Attribute Generation

Following the previous work [26], we employed prompts in a question and an answer format, as shown in Table 9.

Table 9. Question–answer prompts used for generating attributes.

Dataset	Prompts
CIFAR-10 CIFAR-100 ImageNet-1K	**Q**: What are useful features for distinguishing a {*category name*} in a photo? **A**: There are several useful visual features to tell there is a {*category name*} in a photo:
Lacuna-10	**Q**: What are useful features for distinguishing the face of {*category name*} in a photo? **A**: There are several useful visual features that help identify the face of {*category name*} in a photo, especially in celebrity recognition:

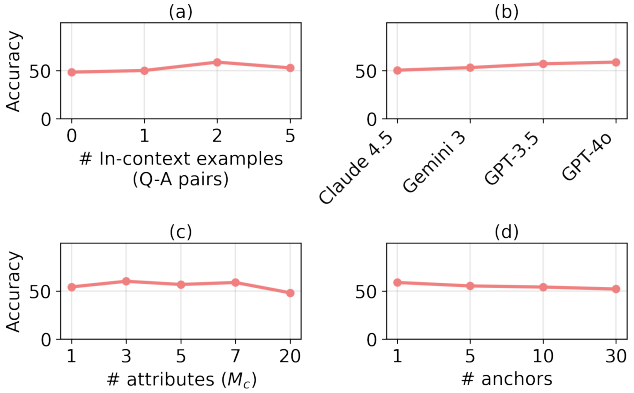

Figure 9. Sensitivity analysis on CIFAR-10 ($k = 256$).

These prompts were crafted to obtain concise and distinctive attribute descriptions for each category, which served as the basis for our semantic anchors. For each dataset, we queried GPT-4o using the question and answer prompts to obtain class-wise attribute descriptions, which are summarized in Table 10. Each prompt instructed GPT-4o to describe the visual or contextual traits of a given category, including elements such as shape, texture, or common environment. Prior work [26] reported that list formatting becomes more reliable when one or two examples are provided, and we included two short examples in each prompt to encourage structured and coherent outputs. The resulting class-wise attributes (shown in Table 10) were then encoded with the semantic encoder to construct class-level anchors. This process produces attribute representations that are semantically coherent and consistently organized, enabling anchor-based structure preservation across datasets.

F. Sensitivity

Since our anchors are derived from semantic priors provided by an external large language model, assessing the stability of the anchor generation process is important. To this end, we conducted a sensitivity analysis on (a) the LLM prompt template, (b) the LLM model, (c) the number of attribute descriptions per class, and (d) the number of anchors per class. The results are summarized in Figure 9, where accuracy denotes the average of A_{test} and A_r. Across all variations, the proposed method consistently achieves stable performance, demonstrating robustness to the choice of anchor generation components. This robustness indicates that semantic anchors provide reliable references for preserving representation structure during unlearning.

Table 10. Examples of attribute descriptions for randomly selected categories.

Dataset	Category	Attributes
CIFAR-10	airplane	“wings extending from the sides”, “tail fin at the back”, “jet engines under the wings or on the tail”, “cockpit windows at the front”, “landing gear”, “airline logos or markings”, “propellers if it’s a propeller plane”
	cat	“furry body”, “pointed ears”, “whiskers”, “tail”, “round eyes, often green or yellow”, “claws”, “variety of colors and patterns, including solid, striped, and spotted”
CIFAR-100	aquarium fish	“fins and tail”, “scales”, “gills”, “small size”, “often found in water or an aquarium setting”, “may be seen swimming with other fish”, “may have unique patterns or markings on their bodies”
	pickup truck	“front cab with two or four doors”, “a large, open cargo area in the back”, “a tailgate at the rear of the cargo area”, “a front grille and headlights”, “side mirrors”, “a license plate”, “a truck bed liner or cover (optional)”
ImageNet-1K	suspension bridge	“a large structure spanning a body of water or other gap”, “typically has two towers or piers supporting the main span”, “the main span is suspended by cables or chains”, “may have a deck for pedestrians, vehicles, or trains”, “may have decorative elements such as lights or flags”
	dust jacket	“a book cover”, “made of paper or cloth”, “has a front and a back”, “usually has a printed design or image”, “may have text on the front and/or back”, “may be brightly colored or have a pattern”
Lacuna-10	Alexis Sánchez	“intense dark brown eyes and closely shaved eyebrows”, “cropped black hair, often styled in a skin fade or taper”, “tan skin with high cheekbones and athletic facial lines”, “clenched jaw or focused game-time expression”, “muscular build, sometimes captured mid-action”, “often seen in athletic wear, particularly a football jersey”
	Alesha Dixon	“almond-shaped dark brown eyes with defined lashes”, “glowing medium-brown skin tone”, “wears bold makeup styles, often with shimmering eye shadow and statement lip colors”, “charismatic smile and dynamic expressions on stage or red carpet”