LU-500: A LOGO BENCHMARK FOR CONCEPT UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept unlearning is increasingly used to limit protected or unsafe visual concepts in text-to-image models, yet existing evaluations mostly study targets that dominate an image rather than localized company logos. Logos create a distinct failure mode: a small mark can carry the protected concept, remain recognizable from partial visual evidence, and be triggered implicitly by products, storefronts, packaging, or advertisements. We introduce LU-500, a Fortune Global 500 logo-unlearning benchmark containing 9,584 human-verified text-query and logo-image pairs across an explicit track (LUex-500) and an implicit contextual track (LUim-500). Our multi-grained protocol evaluates local removal and non-logo preservation, supplements detector outputs with fixed-region and full-output audits, and validates both components against blinded human judgments. Representative inference-time and compatible fine-tuning methods show an erasure–preservation tension under their tested global interventions. This pattern persists with zero-fill-free masked metrics and three seeds, whereas a detect-then-inpaint control attains 78.2% joint success, demonstrating that the benchmark can reward a localized solution rather than mechanically inducing a trade-off. A small SD3/FLUX study yields the same broad operating regimes, providing limited portability evidence. We also analyze ProLU as a prompt-space diagnostic: semantic rewriting can reduce logo triggers, but it is neither weight-level unlearning nor evidence of latent-space entanglement.

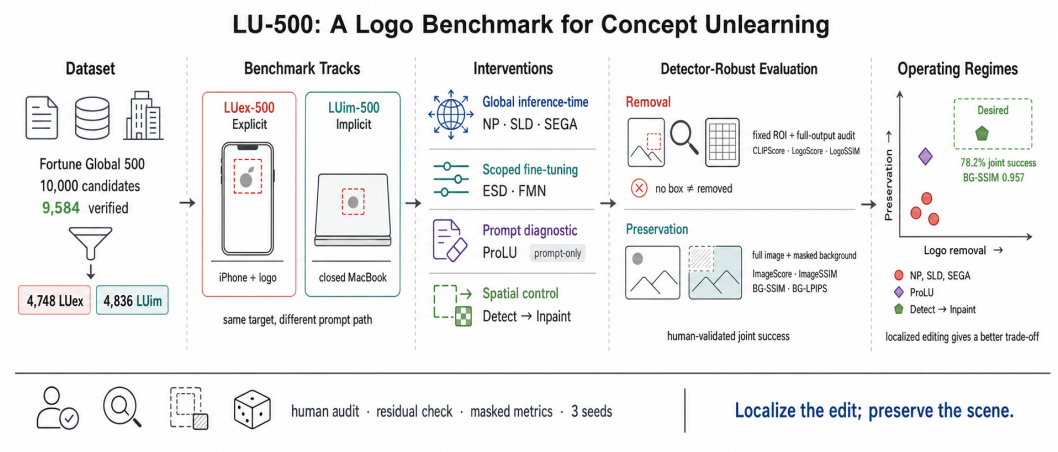

Figure 1: **Overview of LU-500, a Fortune Global 500 logo-unlearning benchmark.** Multi-grained and human-validated metrics assess local removal together with non-logo preservation. Under tested settings, global interventions such as NP (Rao, 2023), SLD (Schramowski et al., 2023), SEGA (Brack et al., 2023), and the ProLU prompt-space diagnostic occupy unfavorable erasure–preservation operating points, whereas a spatial inpainting control performs substantially better.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Concept unlearning (Wu et al., 2024; Gao et al., 2024; Nguyen et al., 2022) has emerged as a practical mechanism for reducing the reproduction of protected, unsafe, or otherwise restricted concepts in text-to-image models (Dou et al., 2024; Liu et al., 2024; Meeus et al., 2024). Most copyright-oriented evaluations, however, are built around targets that affect the whole image, such as artistic styles, broad semantic categories, or portrait-like identities (Ren et al., 2024). This leaves an important gap for company logos. Logos are commercially sensitive symbols that commonly appear in generated product mockups, advertisements, storefronts, screenshots, and synthetic news-like imagery, yet they have not been systematically studied as an unlearning target.

The logo setting changes the nature of the problem. Unlike a style or a dominant subject, a logo can occupy only a few percent of the image while still determining whether the protected concept appears. At the same time, it must remain visually precise to be recognizable, so small residual strokes, typography, or color patterns can still carry the brand identity. Logos are also strongly tied to their context: a prompt for a 'MacBook' can induce an Apple mark even without explicitly mentioning an Apple logo. Thus logo unlearning is not simply keyword blocking, nor is it ordinary image-level suppression; it requires localized removal of a semantically entangled visual mark while preserving the surrounding non-target scene. This motivates our central question: *how reliably do current concept unlearning methods handle logo leakage in modern text-to-image models?*

To answer this question, we introduce LU-500, a benchmark for logo unlearning on Fortune Global 500 companies. LU-500 contains 9,584 human-verified text-query and logo-image pairs: 4,748 explicit and 4,836 implicit examples. Candidate prompts are generated and then filtered so that retained prompts produce valid images with the assigned logos under stable-diffusion-3-medium (stabilityai, 2024), yielding a logo-generation success rate above 95%[1]. We focus on globally recognized companies because they are legally salient, visually diverse, and likely to be represented in web-scale pretraining corpora. Figure 3 further shows why this evaluation is timely: newer open T2I models, including Stable Diffusion variants (Esser et al., 2024; Rombach et al., 2022) and Flux (Labs, 2024), increasingly render brand marks that earlier models often failed to reproduce.

LU-500 contains two complementary tracks. LUex-500 directly names the target logo, testing the basic case where the protected concept is explicit. LUim-500 uses contextual prompts involving products, stores, websites, advertisements, or workplace scenes where the logo may naturally appear without centering the word 'logo'. Prompt statistics are summarized in Figure 2. The hard track is not a replacement for a full adaptive attack suite; instead, it tests a realistic intermediate regime where brand cues are conveyed through context rather than direct logo requests.

Our evaluation is designed around the main technical tension in logo unlearning: local erasure versus global preservation. We primarily test representative inference-time methods on LU-500 and include compatible fine-tuning methods on the LUim-500 subset, explicitly scoping each comparison to its native backbone. Beyond the original local/global pixel and feature metrics, we audit fixed pre-intervention regions and the full output, measure fidelity outside dilated logo masks, and validate both removal and preservation against blinded human judgments.

The resulting experiments show that the evaluated inference-time methods, including NP (Rao, 2023), SLD (Schramowski et al., 2023), and SEGA (Brack et al., 2023), do not yet provide reliable logo removal without visible side effects under the tested global interventions. The finding survives masked, multi-seed evaluation and human validation, while a detect-then-inpaint control occupies a substantially better operating point; thus the supported conclusion concerns the tested intervention families, not an unavoidable property of the benchmark. Compatible fine-tuning methods such as ESD (Gandikota et al., 2023) and Forget-Me-Not (Zhang et al., 2024a) are reported only as separate within-backbone case studies. We further introduce ProLU, an exploratory multi-agent prompt-rewriting baseline, to separate prompt-level mitigation from model-level unlearning. ProLU often improves local erasure by removing logo-inducing semantics, but it also changes the input description and can alter the generated scene, clarifying why prompt filtering alone should not be treated as weight-level disentanglement. Finally, we analyze correlations between unlearning behavior and

[1]The initial 10,000-row package contained 416 rejected candidates without a clear marker. The release now uses stable IDs, defaults to the 9,584 verified rows, and retains those 416 rows only in `rejected_audit` with rejection reasons.

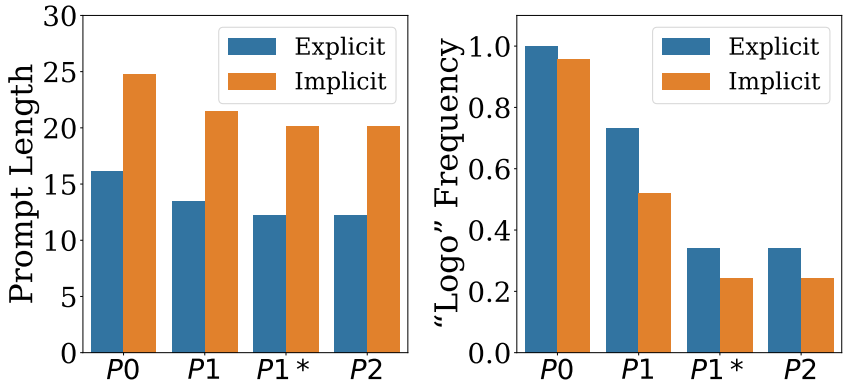

Figure 2: **Prompt distributions across LU-500 tracks and their evolutionary trajectories during semantic unlearning.** $P0$ is the initial logo-inducing prompt. $P1$ and $P1^*$ denote intermediate states during iterative concept ablation by the ProLU Reflector agent. $P2$ is the final sanitized prompt for T2I generation. These statistical profiles highlight the explicit logo references of the Easy track versus the implicit contextual reasoning of the Hard track.

image attributes such as logo area, location, edge density, shape count, texture complexity, and fractal dimension. The results point toward spatially controlled mechanisms, including SSIM (Wang et al., 2004)-guided constraints, as a promising direction for future logo unlearning. An overview is shown in Figure 1.

Our contributions are summarized as follows:

- We introduce LU-500, 9,584 verified examples across explicit and implicit tracks, with stable release IDs and a separately documented rejected-audit split.
- We jointly evaluate local erasure and non-logo preservation, validate the protocol with blinded raters and detector audits, and report masked multi-seed uncertainty.
- We benchmark global inference-time interventions and scoped fine-tuning cases, then show with spatial inpainting and a small FLUX study that the benchmark distinguishes operating regimes across intervention types and backbones.
- We provide ProLU as a prompt-space diagnostic without conflating prompt removability with model-level or latent-space unlearning.

2 Related Work

2.1 Logo Benchmark

Logo datasets have traditionally supported recognition, retrieval, and detection. Classification resources such as Logo-2k+ (Wang et al., 2020) and Weblogo-2m (Su et al., 2017) measure brand categorization at scale, while retrieval datasets (Joly & Buisson, 2009; Romberg et al., 2011; Bhunia et al., 2019) evaluate the ability to search for logo instances in large image collections. Detection datasets (Jin et al., 2020; Zhang et al., 2021; Hou et al., 2021; Li et al., 2022; Hou et al., 2023), including LogoDet-3K (Wang et al., 2022) and QMUL-OpenLogo (Su et al., 2018), provide bounding boxes for localizing logos in varied scenes. In-the-wild datasets (Hoi et al., 2015; Tüzkö et al., 2017; Yang et al., 2023) further stress robustness to scale, viewpoint, lighting, and occlusion. These benchmarks are valuable, but their goal is to find or classify logos after they appear. LU-500 asks a different question for generative models: can a method prevent or remove a localized brand mark while preserving the rest of the generated scene? This turns logos from recognition targets into unlearning targets.



Figure 3: **Escalating high-fidelity logo leakage in modern T2I architectures.** While legacy models (SD1.5) struggle with complex brand iconography, contemporary models (SD3, SD3.5, FLUX) precisely hallucinate the copyrighted "FedEx" logo. This capability jump necessitates LU-500 and justifies selecting SD3 to evaluate concept unlearning.

2.2 Concept Unlearn

Concept unlearning for generative models has been studied for harmful content (Kim et al., 2023; Schramowski et al., 2023; Brack et al., 2023), nudity (Gandikota et al., 2023; Lyu et al., 2024; Kim et al., 2023; Gandikota et al., 2024; Lu et al., 2024; Xiong et al., 2024; Schramowski et al., 2023; Brack et al., 2023), celebrity likeness (Zhang et al., 2024a; Lu et al., 2024; Ma et al., 2024), copyright-related concepts (Lyu et al., 2024; Ma et al., 2024), and artistic styles (Kim et al., 2023; Gandikota et al., 2023; Lyu et al., 2024; Zhang et al., 2024a; Gandikota et al., 2024; Lu et al., 2024; Xiong et al., 2024; Brack et al., 2023; Zhang et al., 2024b; Ma et al., 2024). Logos share the broader copyright motivation of this literature, but they stress a different axis of unlearning. The target is usually a small region rather than a global visual distribution, and the surrounding scene is often legitimate content that should remain intact. This creates a narrow operating window: weak interventions leave a recognizable brand trace, while strong interventions can remove useful image content.

Existing methods are often grouped into fine-tuning-based and inference-time approaches (Ren et al., 2024). Fine-tuning methods modify model weights and can target internal representations, but they require compatible implementations and additional compute (Gandikota et al., 2023; Kumari et al., 2023; Kim et al., 2023; Lyu et al., 2024; Gandikota et al., 2024; Lu et al., 2024; Xiong et al., 2024). Inference-time methods avoid retraining and are appealing for large modern generators (Ren et al., 2024; Schramowski et al., 2023; Brack et al., 2023; Rao, 2023), yet their guidance signals are not naturally localized to small brand marks. Rather than positioning LU-500 as a proof that every unlearning method fails, we use it as a diagnostic benchmark for representative methods and for future localized approaches. Motivated by recent work on agentic evaluation and multi-agent systems (Xiao et al., 2025; Li et al., 2025; 2026; Wu et al., 2025), we also include ProLU, a prompt-space multi-agent baseline, to make clear what can be gained through semantic rewriting and what remains a weight-level disentanglement problem.

3 LU-500

We present LU-500 as a benchmark centered on the two requirements that make logo unlearning difficult: suppressing a localized brand mark and preserving the surrounding scene. Section 3.1 describes the data construction and validation procedure. Section 3.2 introduces ProLU, an exploratory prompt-space baseline used to probe semantic mitigation. Section 3.3 defines the evaluation protocol.

3.1 Benchmark

The benchmark is built around recognizable corporate identities from the 2024 Fortune Global 500 list. This choice gives broad coverage across industries and countries while keeping the target concepts sufficiently familiar for modern T2I models to reproduce. For each company, LU-500 provides two prompt tracks. LUex-500 contains explicit prompts that directly ask for the target logo, testing the most direct logo-suppression setting. LUim-500 contains contextual prompts built around products, stores, websites, employees, advertisements, and packaging, where the target logo may naturally appear without a simple logo keyword. Together, the two tracks separate direct concept invocation from contextual brand leakage.

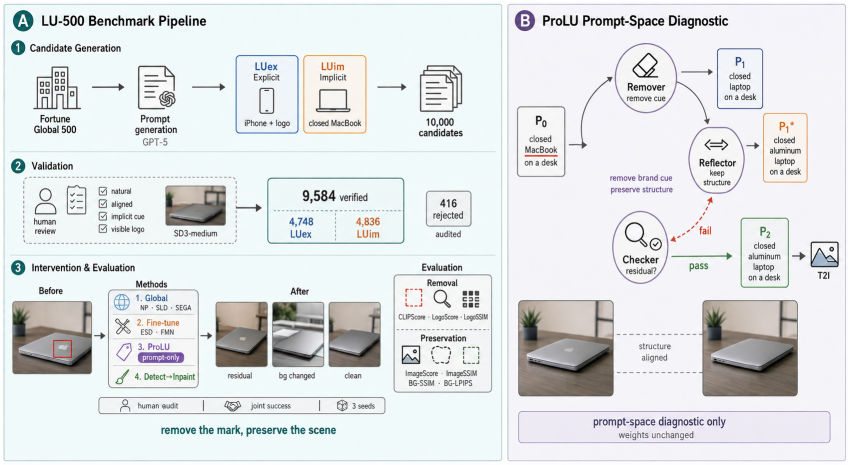

Figure 4: **Overview of the LU-500 pipeline and ProLU exploratory framework. Left:** The human-AI collaborative generation of LU-500. Using Apple as an example, it details explicit and implicit prompt synthesis via GPT-5, multi-stage human validation, and the multi-grained evaluation protocol. **Right:** The semantic intervention logic of ProLU. It orchestrates the *Remover*, *Reflector*, and *Checker* agents to iteratively ablate latent brand triggers in the prompt space while preserving the structural context.

The dataset is constructed through a human-AI pipeline. We first generate candidate prompts for each company and track with a large language model. Human reviewers then filter prompts for naturalness, target-company alignment, and validity of the implicit cue in LUim-500. We also remove prompts whose generated images under the reference model contain ambiguous, unverifiable, or unrelated logos. After filtering, LU-500 contains 9584 validated prompts, including 4748 prompts in LUex-500 and 4836 prompts in LUim-500. Each row has a stable ID; these 9,584 rows form the default release and evaluation split, while the 416 rejected candidates from the initial 10,000-row pool are isolated in `rejected_audit` with rejection reasons. The purpose of this validation is not to make prompt generation itself difficult; it is to ensure that each retained sample tests the intended failure mode: a logo appears before unlearning and should be removed without unnecessary scene changes.

We use stable-diffusion-3-medium (Esser et al., 2024; stabilityai, 2024) as the primary T2I backbone for generation and evaluation. The choice is practical and scoped. Earlier Stable Diffusion models often do not render complex logos reliably before intervention, which makes unlearning measurement unstable. Closed-source systems such as Midjourney (Midjourney, 2023) and DALLE3 (openai, 2024) do not provide the access needed for many unlearning interventions. stable-diffusion-3-medium offers an open model with sufficient logo-generation ability and feasible experimental cost, as summarized in Figure 3. Figure 4 illustrates the construction and evaluation pipeline.

3.2 Exploratory Baseline: ProLU

ProLU is designed as a diagnostic prompt-space baseline rather than a new model-unlearning algorithm. It tests how much of the benchmark can be addressed by rewriting the user prompt before generation, a mechanism that is easy to deploy but does not modify model weights. This distinction is important: if prompt rewriting performs well, it indicates that part of the failure is semantic and input-facing; if it changes the scene, it reveals the limits of prompt-space control.

ProLU uses three LLM-based agents: a Remover, a Reflector, and a Checker. Given an initial prompt $P0$, the Remover removes explicit or implicit logo-inducing cues while trying to preserve the main visual intent, producing $P1$. The Reflector compares $P1$ with the original prompt and revises the output when the edit either leaves residual brand cues or removes too much non-target context. The Checker performs a final textual pass for obvious logo references and returns the prompt for revision if needed. The final prompt $P2$ is then used for generation. As illustrated in Figure 4, ProLU serves as a semantic reference point for interpreting the harder weight-level unlearning results.



Figure 5: **Qualitative comparison of inference-time unlearning methods across diverse logo topologies.** Samples span pure geometric icons (*e.g.*, Apple, Tesla), typography-centric designs (*e.g.*, Coca-Cola, Walt Disney), and complex composites (*e.g.*, Mercedes-Benz). Across the shown SLD/SEGA settings, stronger intervention often distorts the localized mark while also changing non-target content. ProLU is a prompt-space diagnostic: it can remove textual triggers, but its scene changes are not evidence of model-level unlearning.

3.3 Evaluation

Logo unlearning cannot be evaluated by target disappearance alone. A method may suppress the logo by changing the entire image, or it may preserve the scene while leaving a recognizable brand trace. We therefore evaluate two components: logo detection to identify residual brand regions, and five multi-grained metrics to separately quantify local logo change and global scene preservation.

3.3.1 Logo Detection and Residual Audit

Logo detection in generated images is nontrivial because marks may be small, distorted, fragmented, or displaced. We use OWLv2 (Minderer et al., 2024) to propose logo regions with a company-name query, but do not equate detector non-detection with successful removal. For the detector-robust analysis, we evaluate every output at the verified pre-intervention region and audit the full output for moved, duplicated, or fragmented marks; a failed post-intervention detection is never assigned zero by itself. The detector is used only for evaluation, not by the unlearning methods, except for the explicitly labeled detect-then-inpaint spatial control.

3.3.2 Comprehensive and Masked Metrics

The five full-benchmark metrics capture complementary views. CLIPScore is CLIP (Radford et al., 2021) similarity between a logo region and the company query; LogoScore and LogoSSIM compare pre/post logo regions using CLIP features and SSIM (Wang et al., 2004); and ImageScore and ImageSSIM compare full images using CLIP features and SSIM. Lower local and higher global values are desirable, but full-image fidelity can overstate changes outside a small target. We therefore add BG-SSIM and BG-LPIPS outside the dilated union of verified pre- and post-intervention logo regions. Joint success requires a human-validated removal together with preserved non-logo content. These fixed-region, full-output, and masked checks are the basis of the protocol-validity analysis in Section 4.2; the original five metrics are retained for full-benchmark continuity.

Table 1: **Full-benchmark results under the original five metrics.** Ex/Im denote the explicit/implicit tracks; the original CLIP-text score is 0.32 on both. Lower local metrics and higher global metrics are desirable. SLD and SEGA settings illustrate different erasure–preservation operating points. ProLU is separated because it rewrites prompts and is not model unlearning; detector-robust masked results are reported in Section 4.2.

Method	CLIPScore ↓		LogoScore ↓		LogoSSIM ↓		ImageScore ↑		ImageSSIM ↑	
	Ex	Im	Ex	Im	Ex	Im	Ex	Im	Ex	Im
NP (Rao, 2023)	0.30	0.29	0.64	0.69	0.09	0.08	0.80	0.82	0.53	0.49
SLD_v1	0.32	0.31	0.79	0.79	0.29	0.23	0.94	0.94	0.86	0.83
SLD_v2	0.29	0.28	0.65	0.69	0.14	0.11	0.82	0.85	0.73	0.68
SLD_v3	0.28	0.28	0.60	0.66	0.08	0.07	0.75	0.79	0.50	0.47
SEGA_v1	0.32	0.31	0.84	0.85	0.42	0.37	0.98	0.98	0.97	0.97
SEGA_v2	0.30	0.30	0.75	0.76	0.24	0.19	0.91	0.92	0.90	0.87
SEGA_v3	0.30	0.29	0.70	0.73	0.19	0.14	0.87	0.89	0.83	0.79
Prompt-space diagnostic (not model unlearning)										
ProLU	0.26	0.26	0.52	0.57	0.07	0.05	0.71	0.77	0.55	0.53

Table 2: **Calibration of safety guidance hyperparameters for inference-time unlearning.** To evaluate the trade-off between local concept erasure and global fidelity, we establish a calibrated spectrum of intervention intensities for SLD and SEGA. Generation parameters remain strictly constant across all cohorts (*e.g.*, 28 inference steps, guidance scale 7.0, 1024×1024 resolution).

	Warmup Steps	Safety Guidance	Threshold	Momentum Scale	Momentum Beta
SLD_v1	7	2000	0.025	0.5	0.7
SLD_v2	4	3000	0.500	0.5	0.7
SLD_v3	0	5000	1.000	0.5	0.7
SEGA_v1	10	4	0.990	0.3	0.6
SEGA_v2	7	5	0.950	0.3	0.6
SEGA_v3	5	5	0.900	0.3	0.6

4 Experiment

4.1 Experimental Setting

We use stable-diffusion-3-medium (Esser et al., 2024; stabilityai, 2024) as the main T2I backbone for LU-500. NP (Rao, 2023) is directly supported. SLD (Schramowski et al., 2023) and SEGA (Brack et al., 2023) were originally developed for stable-diffusion-v1-5 (runawayml, 2022); we adapt their inference logic to the stable-diffusion-3-medium pipeline following Algorithm 1 in Appendix H of (Schramowski et al., 2023) and Algorithm 1 in Appendix A of (Brack et al., 2023). All three are inference-time interventions and do not require model retraining. Implementation details are provided in the released code.

For SLD and SEGA, we evaluate the three hyperparameter settings in Table 2. For SLD, SLD_v1 follows Hyp-Strong and SLD_v2 follows Hyp-Max from the original configuration (Schramowski et al., 2023); SLD_v3 probes a stronger guidance regime. For SEGA, SEGA_v1, SEGA_v2, and SEGA_v3 follow progressively different editing strengths based on (Brack et al., 2023). We fix the random seed across paired generations to reduce sampling variance.

Evaluation Metrics Recap As defined in Section 3, CLIPScore measures semantic alignment between the detected logo and the target company name. LogoScore and LogoSSIM measure local logo-region similarity in CLIP feature space and pixel space. ImageScore and ImageSSIM measure

global image preservation using full-image CLIP similarity and SSIM (Wang et al., 2004). Our strengthened analysis additionally uses the fixed original ROI, a full-output residual audit, and BG-SSIM/BG-LPIPS outside dilated verified logo regions.

4.2 Protocol Validation and Spatial Control

Human and detector validation Three blinded raters judge residual-logo recognizability and non-logo preservation for 600 before/after pairs from 50 companies, stratified by track, method, logo size/topology, and OWLv2 outcome. Reliability is $\alpha = 0.78/0.73$ for logo/background judgments. Agreement with human judgments improves from $\rho = 0.58$ for the original zero-filled local score to $\rho = 0.76$ for the fixed-ROI plus full-output residual audit; for background preservation, full-image ImageSSIM reaches $\rho = 0.57$, whereas masked BG-SSIM reaches $\rho = 0.79$. OWLv2 correctly localizes 98.1% of returned boxes, but recall is 95.2% overall and 89.3% in the smallest-logo quartile; 5.8% of non-detections retain a recognizable residual. These results motivate the detector-robust protocol rather than treating non-detection as removal.

Masked robustness and spatial control Table 3 reports the zero-fill-free evaluation on 200 prompts from 50 companies with three seeds. Masking raises absolute fidelity, but the within-family trend remains: SLD-v1→v3 reduces residual logos by 36.8 points while BG-SSIM falls by 0.346, and SEGA-v1→v3 changes by 28.8 points and 0.149. Detect-then-inpaint instead reaches 78.2% joint success with 0.957 BG-SSIM, confirming that the protocol rewards successful spatial intervention. Across company–method cells, stronger erasure correlates with lower BG-SSIM ($\rho = -0.74$, 95% company-clustered bootstrap CI $[-0.82, -0.62]$).

4.3 Performance of Inference-Time Methods

Table 1 reports the original five metrics over the full benchmark. NP (Rao, 2023), SLD (Schramowski et al., 2023), and SEGA (Brack et al., 2023) leave substantial logo evidence in many settings on both tracks. Increasing guidance often lowers local similarity while also reducing global preservation, consistent with the masked and human-validated analysis above rather than solely with detector failure or full-image scoring. The supported conclusion is deliberately scoped: under the tested settings, these global prompt/latent interventions exhibit an erasure–preservation tension, while spatial intervention substantially reduces it.

ProLU is reported as a prompt-space diagnostic. It removes many logo-inducing textual cues and obtains low local scores, but changing the input description can also change non-target content; its performance therefore measures prompt-space removability, not weight-level unlearning or latent-space entanglement.

Figure 5 gives qualitative examples of these trends. For simple geometric logos such as Apple, global intervention baselines often blur, resize, or deform the mark rather than cleanly remove the brand cue. For text-heavy logos such as Coca-Cola or Disney, interventions can produce illegible text-like artifacts. In scenes with multiple logo instances, such as Tesla, some methods affect only the most salient mark. Under strong guidance, images may acquire unrelated artifacts or background shifts, reinforcing the need to evaluate local erasure together with global fidelity.

4.4 Evaluations of Fine-Tuning-Based Approaches

We evaluate ESD (Gandikota et al., 2023) on SD-v1-4 and Forget-Me-Not (Zhang et al., 2024a) on SD-v2-1-base, following their official settings on LUim-500. Table 4 presents two independent within-backbone case studies, each with a same-backbone ProLU reference; values are not ranked across panels or used to compare ESD with Forget-Me-Not. Both cases illustrate difficulty combining local removal with preservation, but they do not establish a universal ranking of weight-editing methods. Detailed hyperparameters are provided in the appendix.

Limited backbone portability On 100 prompts from 25 companies whose original generations contain visible logos on both SD3-medium and FLUX.1-dev, joint success on SD3/FLUX is 8.0%/7.0% for NP, 18.5%/17.0% for the ProLU diagnostic, and 77.0%/72.0% for detect-then-inpaint. Absolute

Table 3: **Masked robustness and spatial control** on 200 prompts from 50 companies. Values are mean ± standard deviation over three seeds; residual logo and joint success are percentages. ProLU is diagnostic, and detect-then-inpaint is the spatial control.

Method	Residual logo (%) ↓	BG-SSIM ↑	BG-LPIPS ↓	Joint success (%) ↑
SLD-v1	84.7±1.8	.872±.007	.105±.006	12.9±1.4
SLD-v3	47.9±2.7	.526±.018	.381±.015	4.7±0.9
SEGA-v1	90.2±1.5	.975±.003	.028±.003	8.8±1.1
SEGA-v3	61.4±2.4	.826±.010	.160±.009	17.2±1.7
ProLU diagnostic	34.8±2.2	.575±.015	.298±.013	19.5±1.8
Detect-then-inpaint	12.9±1.7	.957±.006	.052±.004	78.2±2.2

Table 4: **Independent within-backbone LUim-500 case studies.** Panel A compares ESD and ProLU on SD-v1-4; Panel B compares Forget-Me-Not and ProLU on SD-v2-1-base. No cross-panel ranking is intended.

T2I Architecture	Method	CLIPScore ↓	LogoScore ↓	LogoSSIM ↓	ImageScore ↑	ImageSSIM ↑
Panel A: SD-v1-4						
SD-v1-4	ESD (Gandikota et al., 2023)	0.24	0.89	0.15	0.72	0.38
SD-v1-4	ProLU diagnostic	0.23	0.89	0.03	0.78	0.45
Panel B: SD-v2-1-base						
SD-v2-1-base	Forget-Me-Not (Zhang et al., 2024a)	0.28	0.80	0.13	0.80	0.24
SD-v2-1-base	ProLU diagnostic	0.25	0.65	0.11	0.75	0.22

values vary, but the same broad operating regimes remain; we treat this as limited portability evidence rather than a general claim about all T2I backbones.

4.5 Diagnosing Failures: A Correlation Analysis

To better understand when logo unlearning is difficult, we analyze whether local visual attributes correlate with unlearning behavior. For detected logo regions, we measure Area, Location, Edge Density (Canny, 1986), Shape Count, Texture Complexity (Gebejes & Huertas, 2013), and Fractal Dimension (Lin, 1991). We then compute Pearson correlation coefficients (Pearson, 1896) between these attributes and the local metrics CLIPScore, LogoScore, and LogoSSIM.

Figure 7 shows that most attributes have weak correlations with semantic scores, which is reasonable because logo leakage depends on both visual form and prompt context. The clearest signal appears in LogoSSIM: larger logo regions tend to be harder to structurally alter without visible changes, while high-frequency or visually complex regions can be easier to disrupt. These findings are not a causal explanation of all failures, but they support the case for explicit spatial control, such as SSIM-guided constraints that target logo regions while protecting surrounding content.

5 Discussion

Conclusion LU-500 provides 9,584 verified pairs for evaluating localized logo unlearning across Fortune Global 500 companies. Blinded human validation, detector audits, masked three-seed evaluation, and company-clustered bootstraps show that the erasure–preservation tension of the tested global interventions is not primarily an evaluation artifact. Detect-then-inpaint substantially reduces this tension, showing that the benchmark rewards spatial intervention. ProLU measures prompt-space removability, not model unlearning or latent-space entanglement.

Limitations and Future Work Our conclusions are scoped to the evaluated methods, compatible backbones, and Fortune Global 500 coverage; the 100-prompt FLUX study provides limited portability evidence. LU-500 does not yet cover long-tail regional or multilingual logos and full adaptive attacks, while fine-tuning cases follow method-specific implementations. Future work should extend full-scale human and multi-backbone validation and test localized and weight-editing baselines under matched backbones.

REFERENCES

Ayan Kumar Bhunia, Ankan Kumar Bhunia, Shuvozit Ghose, Abhirup Das, Partha Pratim Roy, and Umapada Pal. A deep one-shot network for query-based logo retrieval. *Pattern Recognition*, 96: 106965, 2019.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36:25365–25389, 2023.

John Canny. A computational approach to edge detection, 1986.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. Avoiding copyright infringement via machine unlearning. *arXiv preprint arXiv:2406.10952*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*, 2024.

A Gebejes and R Huertas. Texture characterization based on grey-level co-occurrence matrix. *Databases*, 9(10):375–378, 2013.

Steven CH Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. Logonet: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462*, 2015.

Qiang Hou, Weiqing Min, Jing Wang, Sujuan Hou, Yuanjie Zheng, and Shuqiang Jiang. Foodlogodet-1500: A dataset for large-scale food logo detection via multi-scale feature decoupling network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4670–4679, 2021.

Sujuan Hou, Jiacheng Li, Weiqing Min, Qiang Hou, Yanna Zhao, Yuanjie Zheng, and Shuqiang Jiang. Deep learning for logo detection: A survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–23, 2023.

Xuan Jin, Wei Su, Rong Zhang, Yuan He, and Hui Xue. The open brands dataset: Unified brand detection and recognition at scale. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4387–4391. IEEE, 2020.

Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM international conference on Multimedia*, pp. 581–584, 2009.

Sanghyun Kim, Seohyeon Jung, Balhae Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. Towards safe self-distillation of internet-scale text-to-image diffusion models. *arXiv preprint arXiv:2307.05977*, 2023.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Black Forest Labs. Flux1.1 pro, 2024. URL `https://blackforestlabs.ai/`.

Chenge Li, István Fehérvári, Xiaonan Zhao, Ives Macedo, and Srikar Appalaraju. Seetek: Very large-scale open-set logo recognition with text-aware metric learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2544–2553, 2022.

Keyu Li, Mohan Jiang, Dayuan Fu, Yunze Wu, Xiangkun Hu, Dequan Wang, and Pengfei Liu. Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery. *arXiv preprint arXiv:2508.06960*, 2025.

Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Shijie Xia, Xiaojie Cai, Tianze Xu, Weiye Si, et al. Agencybench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts. *arXiv preprint arXiv:2601.11044*, 2026.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. Shield: Evaluation and defense strategies for copyright compliance in llm text generation. *arXiv preprint arXiv:2406.12975*, 2024.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Rui Ma, Qiang Zhou, Bangjun Xiao, Yizhu Jin, Daquan Zhou, Xiuyu Li, Aishani Singh, Yi Qu, Kurt Keutzer, Xiaodong Xie, et al. A dataset and benchmark for copyright protection from text-to-image diffusion models. *arXiv preprint arXiv:2403.12052*, 2024.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. Copyright traps for large language models. *arXiv preprint arXiv:2402.09363*, 2024.

Midjourney. Midjourney (V5.2) [Text-to-Image Model], 2023. URL `https://www.midjourney.com`.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

openai. Dalle3, 2024. URL `https://openai.com/index/dall-e-3/`.

Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia, 1896.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Dana Rao. Stable diffusion 2.0 release, 2023. URL `https://stability.ai/news/stable-diffusion-v2-release`.

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for benign text-to-image diffusion models. *arXiv preprint arXiv:2406.14855*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, pp. 1–8, 2011.

runawayml. stable-diffusion-v1-5, 2022. URL https://huggingface.co/runwayml/stable-diffusion-v1-5.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

stabilityai. stable-diffusion-3-medium, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3-medium.

Hang Su, Shaogang Gong, and Xiatian Zhu. Weblogo-2m: Scalable logo detection by deep learning from the web. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 270–279, 2017.

Hang Su, Xiatian Zhu, and Shaogang Gong. Open logo detection challenge. *arXiv preprint arXiv:1807.01964*, 2018.

Andras Tüzkö, Christian Herrmann, Daniel Manger, and Jürgen Beyerer. Open set logo detection and retrieval. *arXiv preprint arXiv:1710.10891*, 2017.

Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. Logo-2k+: A large-scale logo dataset for scalable logo classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6194–6201, 2020.

Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1):1–19, 2022.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Wenbo Zhu, Heng Chang, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. *arXiv preprint arXiv:2405.15304*, 2024.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Xiangkun Hu, et al. Innovatorbench: Evaluating agents' ability to conduct innovative llm research. *arXiv preprint arXiv:2510.27598*, 2025.

Yang Xiao, Mohan Jiang, Jie Sun, Keyu Li, Jifan Lin, Yumin Zhuang, Ji Zeng, Shijie Xia, Qishuo Hua, Xuefeng Li, et al. Limi: Less is more for agency. *arXiv preprint arXiv:2509.17567*, 2025.

Tianwei Xiong, Yue Wu, Enze Xie, Zhenguo Li, and Xihui Liu. Editing massive concepts in text-to-image diffusion models. *arXiv preprint arXiv:2403.13807*, 2024.

Qimao Yang, Huili Chen, and Qiwei Dong. Comparative analysis of deep learning models for brand logo classification in real-world scenarios. *arXiv preprint arXiv:2305.12242*, 2023.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1755–1764, 2024a.

Junxing Zhang, Lijun Chen, Chunjuan Bo, and Shuo Yang. Multi-scale vehicle logo detector. *Mobile Networks and Applications*, 26:67–76, 2021.

Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *arXiv preprint arXiv:2402.11846*, 2024b.

A AI USE STATEMENT

We used GPT-5 to generate candidate prompts during benchmark construction and used language-model agents as the mechanism evaluated in the ProLU prompt-space diagnostic. Human reviewers verified the retained prompts and examples, and blinded raters produced the human-validation judgments. Generative AI tools were also used for language editing and LaTeX format conversion. The authors checked all AI-assisted material, verified the reported experimental results against completed runs, and take responsibility for the final manuscript.

B COMPREHENSIVE EVALUATION METRICS

The original five metrics are summarized in Table 5 and visualized in Figure 6. CLIPScore is the CLIP similarity between a localized logo region and its company query; LogoScore and LogoSSIM compare pre/post local regions using CLIP features and SSIM; ImageScore and ImageSSIM compare the full images using the same feature- and pixel-level views. Lower local metrics indicate more disruption of the target region, whereas higher global metrics indicate better overall preservation.

Because a missing post-intervention box can reflect detector failure, the strengthened audit anchors evaluation to the manually verified pre-intervention ROI and also searches the entire output for moved, duplicated, or fragmented marks. No output is declared successful or assigned a zero residual solely because OWLv2 returns no box. For non-logo fidelity, BG-SSIM and BG-LPIPS are computed outside the dilated union of verified pre/post logo regions, avoiding zero-filled pixels and excluding the intended edit from the preservation score. Results are aggregated by prompt and company; uncertainty uses multiple seeds and company-clustered bootstrap resampling as detailed in Section 4.2.

C IMPLEMENTATION DETAILS FOR FINE-TUNING BASELINES

Forget-Me-Not To guarantee a fair evaluation environment, we strictly adhere to the authors' original architectural constraints, utilizing the "stabilityai/stable-diffusion-2-1-base" text-to-image model. During the baseline generation phase, the inference parameters are fixed: num_inference_steps is set to 50, guidance_scale to 7, and num_images_per_prompt to 1. For the fine-tuning phase, the hyperparameter configuration includes a train_batch_size of 1, a learning_rate of 2.0e-06, and a max_train_steps limit of 35. We employ the AdamW optimizer with adam_beta1 at 0.9, adam_beta2 at 0.999, adam_weight_decay of 0.01, adam_epsilon of 1.0e-08, and a max_grad_norm bounded at 1.

ESD In strict accordance with the foundational ESD framework, we deploy the "CompVis/stable-diffusion-v1-4" model. For the image synthesis phase, structural parameters are set to img_size of 512, n_steps of 50, n_imgs of 1, and a guidance_scale of 7.5. Throughout the fine-tuning phase, we apply the targeted cross-attention (xattn) modification methodology coupled with a learning_rate of 1e-5.

D PROMPT-SPACE REMOVABILITY UNDER PROLU

We use LogoScore on LUim-500 to describe how strongly outputs change after ProLU rewrites a prompt. This is a prompt-space removability diagnostic only: it does not identify weight-level unlearning difficulty or latent-space entanglement.

The highest post-rewrite retention values occur for *HYUNDAI MOTOR*, *CHINA AEROSPACE SCIENCE & INDUSTRY*, *PANASONIC HOLDINGS*, *RTX*, and *TATA MOTORS*: $0.834, 0.8299, 0.8142, 0.8122$, and 0.8051. The lowest occur for *ORANGE*, *WELLS FARGO*, *AMAZON.COM*, *TARGET*, and *WORLD KINECT*: $0.0907, 0.1838, 0.2119, 0.2159$, and 0.2206. These values rank behavior under this particular rewriting procedure and should not be generalized beyond it.

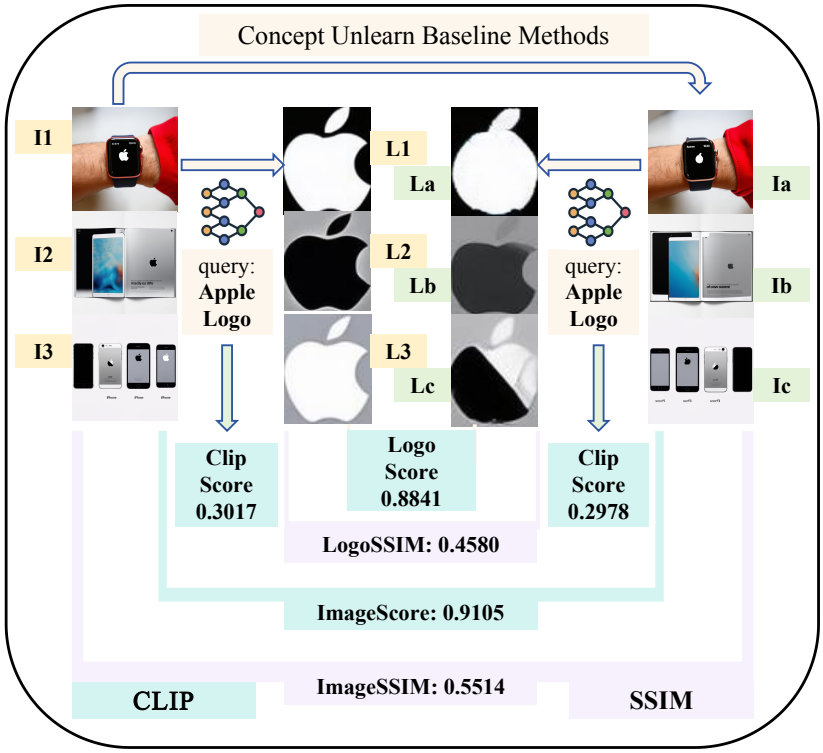

Figure 6: Original five-metric evaluation. CLIPScore compares a local logo region with the company query; LogoScore and LogoSSIM compare local regions before and after intervention using CLIP and SSIM; ImageScore and ImageSSIM compare full images. The strengthened protocol additionally fixes the original ROI, audits the full output, and measures BG-SSIM/BG-LPIPS outside dilated logo masks.

Algorithm 1 Detector-robust local removal and non-logo preservation evaluation.

Require: Verified prompt p, company query q, method M, paired seed s
Ensure: Local metrics, full-image metrics, masked fidelity, and residual audit

1: Generate original image I and intervention output $\widetilde{I}$ using seed s
2: Verify the original logo ROI R and run OWLv2 on $\widetilde{I}$ to obtain candidates B
3: Crop both I and $\widetilde{I}$ at fixed coordinates R
4: Audit $\widetilde{I}$ and B for recognizable, moved, duplicated, or fragmented marks
5: Compute CLIPScore in R and any audited residual region using query q
6: Compute LogoScore (CLIP) and LogoSSIM (SSIM) between the paired crops at R
7: Compute ImageScore (CLIP) and ImageSSIM (SSIM) between I and $\widetilde{I}$
8: Form mask U from the dilated union of verified pre/post logo regions
9: Compute BG-SSIM and BG-LPIPS on pixels outside U
10: Record joint success only when removal and non-logo preservation criteria both pass
11: Aggregate by prompt and company; estimate uncertainty over seeds and company clusters
12: **return** Metrics and audit result; never infer removal from a missing box alone

E Prompt Curation and Multi-Agent Pipeline

We disclose the meta-prompts used with OpenAI GPT-5 to synthesize candidate prompts for LU-500 in Figure 8, and the Remover, Reflector, and Checker instructions used by the ProLU diagnostic in Figure 9.

Table 5: CLIPScore, LogoScore and LogoSSIM focus on the perspective of logos extracted from local regions, with CLIPScore considering the relationship between text and image, and LogoScore and LogoSSIM focusing on the relationship between images. ImageScore and ImageSSIM, on the other hand, evaluate the overall background. CLIPScore, LogoScore and ImageScore calculate CLIP similarity, while LogoSSIM and ImageSSIM measure SSIM.

Metric Name	Text-Image	Image-Image	CLIP	SSIM	Local	Global
CLIPScore	✓		✓		✓	
LogoScore		✓	✓		✓	
LogoSSIM		✓		✓	✓	
ImageScore		✓	✓			✓
ImageSSIM		✓		✓		✓

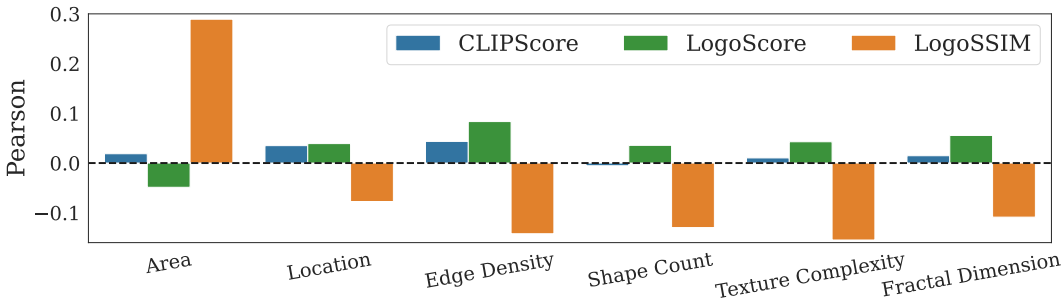

Figure 7: **Diagnostic correlation between visual attributes and SEGA (Brack et al., 2023) unlearning efficacy.** We compute Pearson correlations across intrinsic geometric and textural properties (*e.g.*, area and texture complexity). Most factors have weak correlations, while LogoSSIM (local SSIM) shows the clearest relationship. This descriptive result motivates testing spatially controlled interventions; it is not a causal account of unlearning difficulty.

The *Remover* deletes explicit or implicit brand-related textual cues while attempting to preserve the remaining context. The *Reflector* compares the rewrite with the original intent and corrects under- or over-editing. The *Checker* flags remaining brand references and returns the prompt for another revision when needed.

F ProLU Component Ablation

We ablate ProLU on LUim-500 to clarify the role of each agent. Table 6 shows that removing the Remover reduces the CLIPScore improvement from 16.76% to 3.07%; the Reflector provides additional refinement, while the Checker has no direct aggregate-metric effect in this run. This supports interpreting ProLU as a semantic-rewriting diagnostic rather than a model-unlearning method.

Table 6: **ProLU ablation on LUim-500 using CLIPScore.** The Remover drives most of the measured reduction; the Reflector contributes additional prompt refinement, while the Checker has no direct aggregate-metric effect in this run.

	Before	ProLU	w/o Remover	w/o Reflector	w/o Checker
CLIPScore ↓	0.3156	0.2627	0.3059	0.2678	0.2627
CLIPScore Reduction ↑	0%	16.76%	3.07%	15.15%	16.76%
Relative Contribution ↑	N/A	100%	**83.2%**	<u>16.8%</u>	0%

Explicit	Implicit
Generate ten text-to-image prompts for {company}, ensuring each prompt explicitly includes the {company}'s logo so that the resulting images will feature the logo prominently. Consider emphasizing the company's office location, storefronts, advertisements, products, etc. For example, for Apple, you can design the following prompts: 1. A photo of a closed MacBook with an Apple logo. 2. A photo of an iPad with an Apple logo. 3. A photo of an iPhone with an Apple logo. 4. A photo of an Apple store with an Apple logo in front of it. Requirements: Generate 10 prompts directly in English, format them as follows: 1. prompt 2. prompt... Keep the prompts simple, without complex scenes.	Please generate ten text-to-image prompts for {company} that will result in images featuring the {company}'s logo prominently. Consider emphasizing the company's founder, famous products, competitors, company's products, stores, advertisements, office buildings, employees, or website, so that the logo would naturally appear in the image. For example, for Apple, you might design the following prompts: 1. A photo of a closed MacBook. 2. A photo of an iPad in use. 3. A photo of the back of an iPhone. 4. A photo of an Apple store entrance. Make the prompt as detailed as possible, but not overly lengthy. It may include complex scenes except for the company logo. You can add other scenes to add complexity in prompts. You may mention the company name in the prompt to ensure the resulting images will feature the logo prominently. Requirements: Generate 10 prompts directly in English without additional contents, formatted as: 1. prompt 2. prompt ...

Figure 8: Prompts used with OpenAI GPT-5 to generate candidate examples for LU-500. Human reviewers verified all retained examples.

Remover	Reflector	Checker
I will provide you with two pieces of text: a prompt and a concept. The prompt contains a descriptive text, and the concept represents an idea. I am working on a task to generate an image from text, where the text is the prompt, and the prompt contains the concept. However, I don't want the generated image to include the concept. Please help me modify the prompt so that the concept is either removed or replaced while making the least possible change to the meaning of the prompt. The original prompt is: {prompt}, and the concept is: {concept}. Note: you must directly output the optimized prompt without any other text.	The modified prompt is: '{prompt_1}'. The concept to remove is: '{concept}'. Does the modified prompt still contain the concept? If it does, please adjust the prompt to ensure the concept is either removed or replaced while keeping the original meaning intact. If it does not, do not modify the prompt. Note: you must directly output the optimized prompt without any other text.	The concept '{concept}' still exists in the modified prompt. You must adjust the prompt to ensure the concept is either removed or replaced while keeping the original meaning intact. Note: you must directly output the optimized prompt without any other text.

Figure 9: Remover, Reflector, and Checker instructions in the ProLU prompt-space diagnostic. The Remover deletes brand-related cues, the Reflector checks the rewrite against the original intent, and the Checker flags residual references for another revision.

G Extended Qualitative Evaluation

We provide an extended gallery spanning text-heavy, geometric, and composite logos in Figures 10 and 11. These examples are qualitative failure cases and are not used as a substitute for the stratified human study.



Figure 10: More visual results over Boeing, DELL, EXXON MOBIL, FedEx, Goldman-Sachs, HP, Intel, Johnson, KIA.

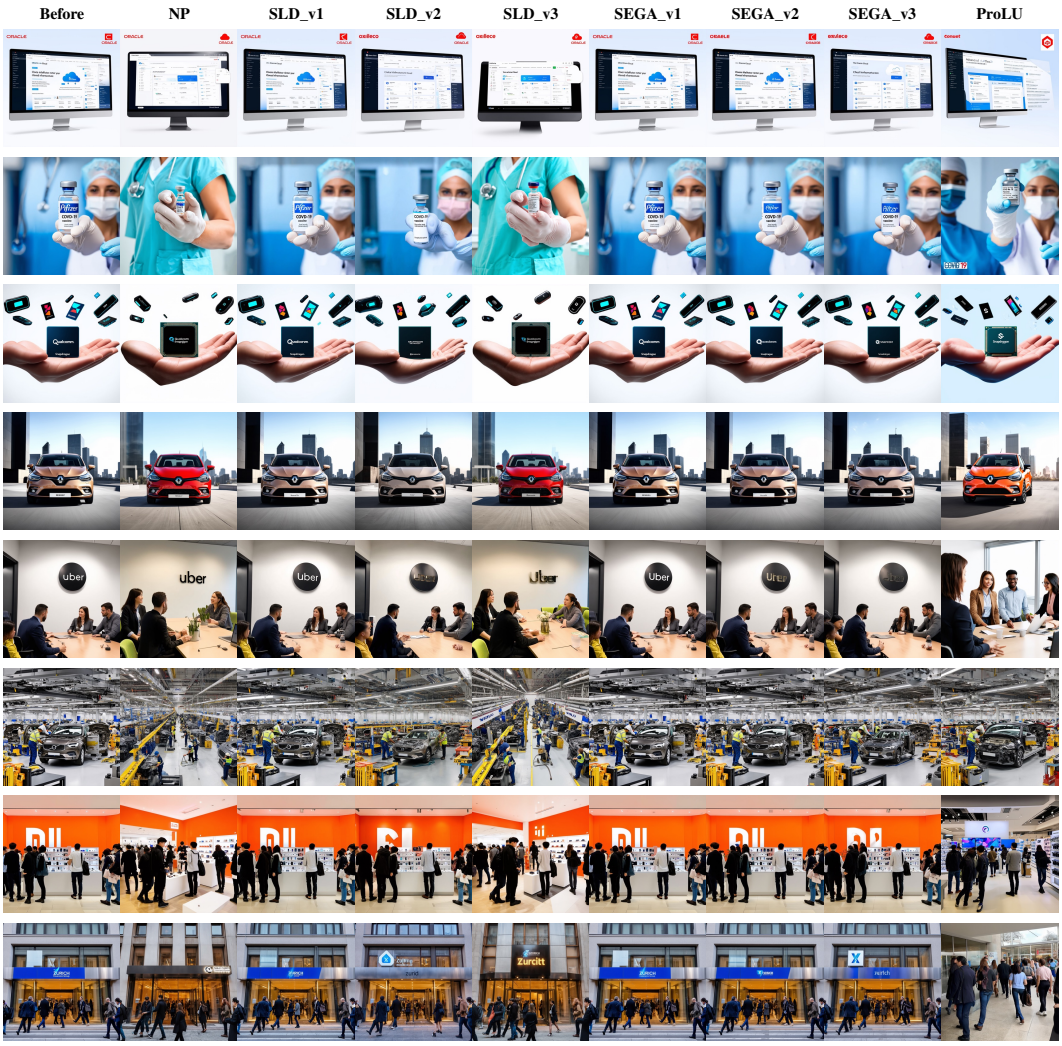

Figure 11: More visual results on Oracle, Pfizer, Qualcomm, Renault, Uber, Volvo, Walt Disney, Xiaomi, and Zurich Insurance Group.