LU-500: A Logo Benchmark for Concept Unlearning

Keyu Li
Shanghai Jiao Tong University
Shanghai, China
chlorophyll@sjtu.edu.cn

Jin Gao
Shanghai Jiao Tong University
Shanghai, China
gaojin@sjtu.edu.cn

Jialing Zhang
Shanghai Jiao Tong University
Shanghai, China
jialingzhang@sjtu.edu.cn

Dequan Wang*
Shanghai Jiao Tong University
Shanghai, China
dequanwang@sjtu.edu.cn

arXiv:2607.24101v2 [cs.CV] 29 Jul 2026

Abstract

Concept unlearning is increasingly used to limit the reproduction of protected or unsafe visual concepts in text-to-image models. Existing evaluations, however, mostly study targets that dominate the whole image, such as styles, broad object categories, or portrait-like identities, leaving company logos comparatively underexamined. Logos create a different failure mode: a small localized mark can carry the entire protected concept, must be visually precise to remain recognizable, and can be triggered implicitly by products, storefronts, packaging, or advertisements even when the word "logo" is absent. We introduce LU-500, a logo-unlearning benchmark built from Fortune Global 500 companies to study this localized and semantically entangled setting. LU-500 contains nearly 10,000 curated text-query and logo-image pairs, with an explicit track (LUex-500) and an implicit contextual track (LUim-500). To avoid reducing the task to a binary detector score, we define a multi-grained protocol that evaluates both local logo removal and global image preservation in pixel and latent spaces. Experiments on representative inference-time methods, including NP, SLD, and SEGA, and compatible fine-tuning-based methods such as ESD and Forget-Me-Not, show that the evaluated methods struggle to remove logo evidence without changing non-target content. We further analyze ProLU, a prompt-space multi-agent baseline: it improves local erasure by removing logo-inducing semantics, but also illustrates why prompt filtering is not a substitute for weight-level disentanglement. Correlation analyses over logo area, location, and structural complexity suggest that future logo unlearning may need spatially aware controls, such as SSIM-guided constraints, rather than purely global concept suppression.

1 Introduction

Concept unlearning [4–6] has emerged as a practical mechanism for reducing the reproduction of protected, unsafe, or otherwise restricted concepts in text-to-image models [7–9]. 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 [10]. 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.

*Corresponding author.
LU-500 is available at GitHub and Hugging Face.

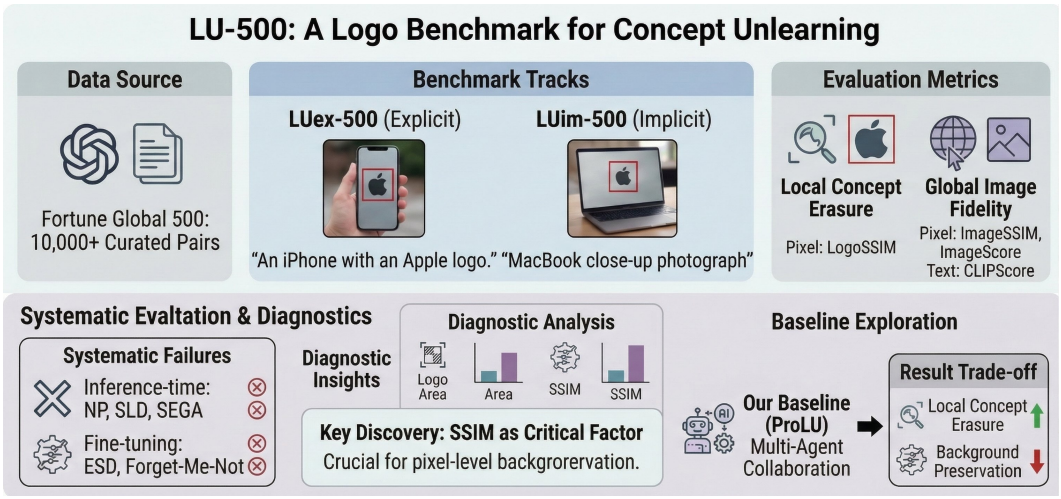

Figure 1: **Overview of LU-500, a benchmark evaluating logo unlearning on the Fortune Global 500.** We propose multi-grained metrics assessing local concept erasure and global image fidelity. Evaluations reveal existing inference-time approaches (*e.g.*, NP [1], SLD [2], SEGA [3]) systematically fail to unlearn these entangled assets. Our prompt-based baseline, ProLU, achieves superior local erasure but exposes a critical trade-off with background preservation. Finally, we diagnose these failures via correlation analysis of intrinsic logo characteristics.

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 nearly 10,000 curated text-query and logo-image pairs. Candidate prompts are generated and then filtered through human verification so that retained prompts produce valid images with the assigned logos under stable-diffusion-3-medium [11], yielding a logo-generation success rate above 95%[2]. 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 [12, 13] and Flux [14], 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-based methods on the LUim-500 subset, explicitly scoping conclusions to these evaluated settings because method compatibility varies across backbones. Rather than using only a binary success rate, we define multi-grained metrics over local and global regions, as well as pixel and latent similarities. These metrics ask whether a method removes the target mark and whether it preserves the rest of the generated image.

[2]We initially constructed 10,000 prompts and filtered out those failing to generate verifiable logos.

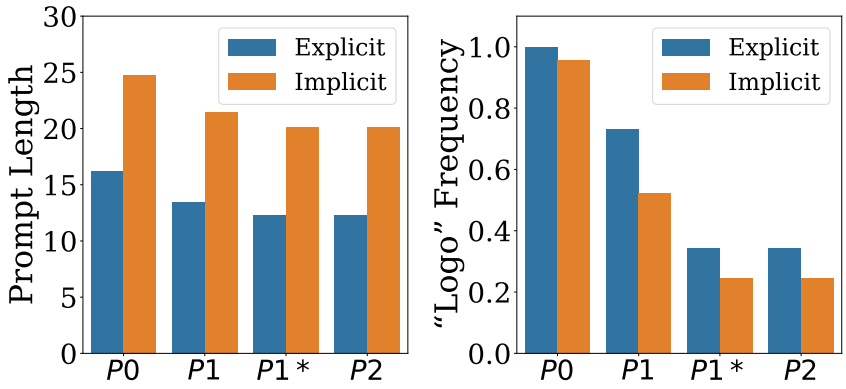

Figure 2: **Prompt distributions across LU-500 tracks and their evolutionary trajectories during semantic unlearning.** $P0$ is the initial logo-inducing prompt. $P1$ and $P1^*$ denote intermediate states during iterative concept ablation by the ProLU Reflector agent. $P2$ is the final sanitized prompt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

The resulting experiments show that the evaluated inference-time methods, including NP [1], SLD [2], and SEGA [3], do not yet provide reliable logo removal without visible side effects. Compatible fine-tuning-based methods such as ESD [15] and Forget-Me-Not [16] show related limitations under their native legacy backbones. 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 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 [17]-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, a benchmark of nearly 10,000 curated text-image pairs for evaluating logo unlearning on Fortune Global 500 companies across explicit and implicit prompt tracks.
- We define a multi-grained evaluation protocol that jointly measures local logo erasure and global image preservation, making the core trade-off measurable rather than hidden inside a single success rate.
- We benchmark representative inference-time and compatible fine-tuning-based unlearning methods, showing that the evaluated methods struggle to remove localized brand marks without changing non-target content.
- We provide ProLU as a diagnostic prompt-space baseline and analyze why semantic rewriting helps in some cases but does not solve localized model-level unlearning.

2 Related Work

2.1 Logo Benchmark

Logo datasets have traditionally supported recognition, retrieval, and detection. Classification resources such as Logo-2k+ [18] and Weblogo-2m [19] measure brand categorization at scale, while retrieval datasets [20–22] evaluate the ability to search for logo instances in large image collections. Detection datasets [23–27], including LogoDet-3K [28] and QMUL-OpenLogo [29], provide bounding boxes for localizing logos in varied scenes. In-the-wild datasets [30–32] 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

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.

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.

2.2 Concept Unlearn

Concept unlearning for generative models [33–35] has been studied for harmful content [36, 2, 3], nudity [15, 37, 36, 38–40, 2, 3], celebrity likeness [16, 39, 41], copyright-related concepts [37, 41], and artistic styles [36, 15, 37, 16, 38–40, 3, 42, 41]. 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 [10]. Fine-tuning methods modify model weights and can target internal representations, but they require compatible implementations and additional compute [15, 43, 36–40]. Inference-time methods avoid retraining and are appealing for large modern generators [10, 2, 3, 1], 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 [44–47], 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.

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

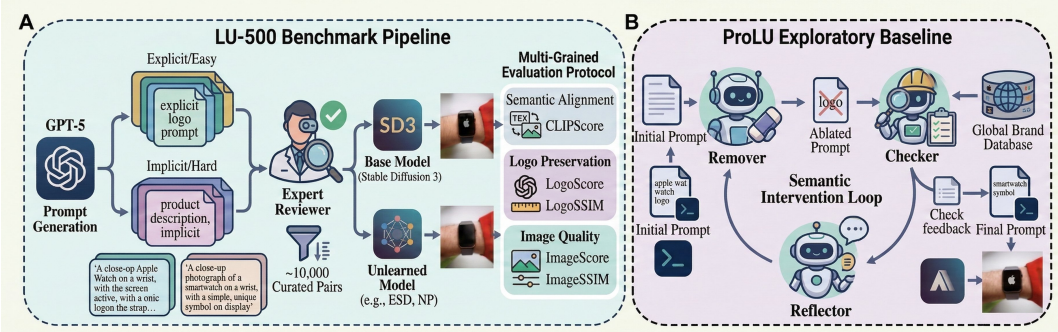

Figure 4: **Overview of the LU-500 pipeline and ProLU exploratory framework. Left:** The human-AI collaborative generation of LU-500. Using Apple as an example, it details explicit and implicit prompt synthesis via GPT-5, multi-stage human validation, and the multi-grained evaluation protocol. **Right:** 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.

logos. After filtering, LU-500 contains 9584 validated prompts, including 4748 prompts in LUex-500 and 4836 prompts in LUim-500. 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 [12, 11] 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 [48] and DALLE3 [49] 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.

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

Logo detection in generated images is nontrivial because the marks may be distorted, partially rendered, or embedded in complex scenes. Supervised detectors can be tied to fixed training

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). These results visually corroborate the systemic trade-off in latent-space interventions: scaling safety guidance (left to right for SLD/SEGA) distorts the localized logo but catastrophically degrades global image fidelity. While our semantic baseline (ProLU) achieves superior erasure, it underscores the inescapable challenge of preserving highly entangled backgrounds.

distributions, while template-based matching can be brittle under scale, pose, and typography changes. We use OWLv2 [50], an open-vocabulary detector, to localize candidate logo regions with text queries such as a company name followed by 'logo'. To reduce evaluator noise, we manually verify the automatically produced boxes on the generated corpus and observe 98% detection accuracy in this verification. The detector is used only for evaluation, not for the unlearning methods.

3.3.2 Comprehensive Metrics

Many inference-time baselines [2, 3, 1] are often summarized with binary success rates. For logos, this is too coarse because the desired edit is spatially small but semantically important. We define five metrics that jointly capture local erasure and global preservation. When no logo is detected after unlearning, local metrics are assigned zero, corresponding to no measurable residual logo region under the detector.

CLIPScore This metric measures CLIP [51] cosine similarity between the detected logo region and the corresponding company-name query. Lower scores indicate weaker semantic alignment with the target brand.

LogoScore and LogoSSIM These metrics compare the detected logo region before and after unlearning. LogoScore uses CLIP image-feature similarity, while LogoSSIM uses pixel-level Structural Similarity (SSIM) [17]. Lower values indicate stronger local disruption of the target logo region.

ImageScore and ImageSSIM These metrics evaluate whether the non-target scene is preserved by comparing the full image before and after unlearning. ImageScore measures global CLIP image similarity, and ImageSSIM measures global SSIM. Higher values are desirable, indicating that logo removal has not unnecessarily degraded the broader image content.

Table 1: **Quantitative evaluation reveals inference-time unlearning's systemic inefficacy.** Original CLIP-Text scores 0.32 on explicit (Ex) and implicit (Im) tracks. Best and second-best are **bold** and underlined. Stronger safety guidance in SLD [2] and SEGA [3] marginally improves local erasure (lower CLIPScore-LogoSSIM) but catastrophically degrades global fidelity (plummeting ImageScore, ImageSSIM). Our baseline (ProLU) achieves deep local ablation but lacks global preservation, proving the inescapable trade-off in disentangling corporate assets.

Method	CLIPScore ↓		LogoScore ↓		LogoSSIM ↓		ImageScore ↑		ImageSSIM ↑	
	Ex	Im	Ex	Im	Ex	Im	Ex	Im	Ex	Im
NP [1]	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
ProLU	**0.26**	**0.26**	**0.52**	**0.57**	**0.07**	**0.05**	0.71	0.77	0.55	0.53

4 Experiment

4.1 Experimental Setting

We use stable-diffusion-3-medium [12, 11] as the main T2I backbone for LU-500. NP [1] is directly supported. SLD [2] and SEGA [3] were originally developed for stable-diffusion-v1-5 [52]; we adapt their inference logic to the stable-diffusion-3-medium pipeline following Algorithm 1 in Appendix H of [2] and Algorithm 1 in Appendix A of [3]. 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 [2]; SLD_v3 probes a stronger guidance regime. For SEGA, SEGA_v1, SEGA_v2, and SEGA_v3 follow progressively different editing strengths based on [3]. 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 [17].

4.2 Performance of Inference-Time Methods

Table 1 shows a consistent local-global trade-off for the evaluated inference-time methods. NP [1], SLD [2], and SEGA [3] reduce local logo metrics only partially in many settings, and CLIPScore remains well above the ideal value of zero. The pattern appears in both explicit and implicit tracks, indicating that logo leakage is not fully addressed by suppressing direct logo keywords.

Increasing safety or editing guidance often improves CLIPScore, LogoScore, and LogoSSIM, but the same settings reduce ImageScore and ImageSSIM. In other words, stronger guidance can disrupt the target region, but it also changes global content, texture, layout, or object identity. This is precisely the failure mode that LU-500 is designed to reveal: for logos, a useful intervention must be local enough to preserve the image while still removing the protected mark.

ProLU provides a complementary reference point. Because it operates by rewriting prompts, it can remove many logo-inducing semantic cues and achieves the strongest local erasure scores in Table 1.

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

However, the improvement is obtained by changing the input description, so non-target scene content can also change. This result is useful diagnostically: prompt-space filtering can mitigate some explicit and contextual triggers, but it does not solve model-level logo unlearning.

Figure 5 gives qualitative examples of these trends. For simple geometric logos such as Apple, latent 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.

Table 3: **Evaluating fine-tuning unlearning architectures on LUim-500.** Best results are **bold**. For fairness, our baseline (ProLU) uses each method's legacy checkpoint. ProLU outperforms ESD and Forget-Me-Not in local erasure (lower CLIPScore-LogoSSIM) without expensive weight updates. Conversely, Forget-Me-Not resists local unlearning but preserves higher global fidelity (ImageScore, ImageSSIM), confirming the inescapable systemic trade-off.

T2I Architecture	Method	CLIPScore ↓	LogoScore ↓	LogoSSIM ↓	ImageScore ↑	ImageSSIM ↑
SD-v1-4	ESD [15]	0.24	0.89	0.15	0.72	0.38
SD-v1-4	ProLU	**0.23**	**0.89**	**0.03**	**0.78**	**0.45**
SD-v2-1-base	Forget-Me-Not [16]	0.28	0.80	0.13	**0.80**	**0.24**
SD-v2-1-base	ProLU	**0.25**	**0.65**	**0.11**	0.75	0.22

4.3 Evaluations of Fine-Tuning-Based Approaches

We also evaluate two fine-tuning-based methods, ESD [15] and Forget-Me-Not [16], on the LUim-500 subset. This comparison is constrained by implementation compatibility: ESD is evaluated on SD-v1-4 and Forget-Me-Not on SD-v2-1-base, following their official settings. ProLU is evaluated separately within each corresponding backbone as a prompt-space reference.

Table 3 shows that these compatible fine-tuning settings also do not consistently achieve both strong local logo removal and high global preservation. We therefore treat the result as scoped evidence rather than a universal statement about all weight-editing methods or modern backbones. Still, the same tension appears: more disruptive behavior can reduce logo evidence but may also lower background fidelity, while conservative behavior preserves the image at the risk of leaving brand traces. Detailed fine-tuning hyperparameters are provided in the supplementary materials.

4.4 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 [53], Shape Count, Texture Complexity [54], and Fractal Dimension [55]. We then compute Pearson correlation coefficients [56] between these attributes and the local metrics CLIPScore, LogoScore, and LogoSSIM.

Table 4: **Quantitative ablation of the ProLU framework on LUim-500.** Measured via CLIPScore, we isolate the exact contribution of each agent. The primary semantic ablator (Remover) drives over 80% of the unlearning efficacy. The semantic stabilizer (Reflector) provides critical refinement, contributing ∼17%. While the deterministic safeguard (Checker) yields a 0% direct metric delta, it remains architecturally essential to guarantee the zero-tolerance robustness of the revision loop.

	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%

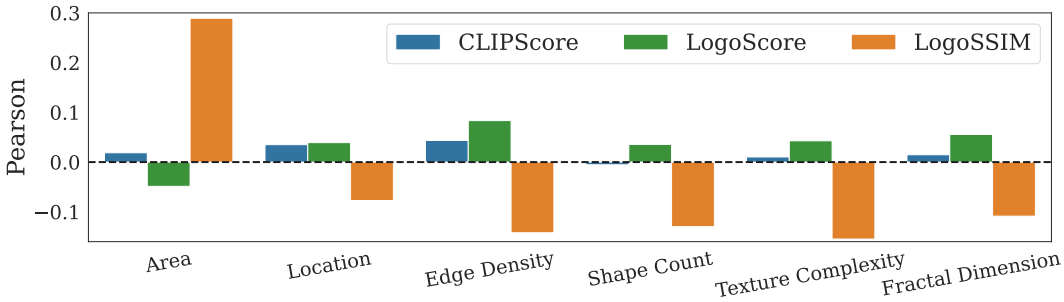

Figure 6: **Diagnostic correlation between visual attributes and SEGA [3] unlearning efficacy.** We compute Pearson correlations across intrinsic geometric and textural properties (*e.g.*, area, texture complexity). While most factors exhibit negligible correlation, LogoSSIM (local SSIM) reveals a notable relationship. This suggests that incorporating structural heuristics (*e.g.*, SSIM-guided spatial controls) is vital for future latent-space unlearning architectures.

Figure 6 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.

4.5 Ablation of the Exploratory Baseline

We ablate ProLU on LUim-500 to clarify the role of each agent. Table 4 shows that the Remover is the main contributor: without it, the CLIPScore reduction drops from 16.76% to 3.07%. The Reflector provides additional stabilization by revising prompts that are either under-edited or over-edited. The Checker has little direct effect on the aggregate metric in this experiment, but it provides a final textual guard against obvious residual logo references. Overall, the ablation confirms that ProLU's gains mainly come from semantic rewriting, reinforcing its role as a diagnostic prompt-space baseline rather than a weight-level unlearning method.

5 Discussion

Conclusion We introduced LU-500, a benchmark of nearly 10,000 text-image pairs for studying logo unlearning on Fortune Global 500 companies. The benchmark frames logos as localized but semantically entangled concepts: removing them requires changing a small protected mark while preserving the rest of the generated scene. Across the evaluated inference-time and compatible fine-tuning-based methods, our multi-grained metrics reveal a persistent local-global trade-off that single success rates would obscure. ProLU further shows that prompt-space rewriting can reduce logo triggers, but also clarifies that prompt filtering is not the same as localized model-level unlearning.

Limitations and Future Work Our empirical conclusions are scoped to the evaluated methods, compatible backbones, and Fortune Global 500 coverage. LU-500 does not yet cover long-tail regional brands, multilingual logo variants, or full adaptive attack protocols, and the fine-tuning comparisons remain constrained by each method's official implementation. Future work should extend the benchmark along these axes, add stronger and more diverse weight-editing baselines where compatible, and report broader variance analyses.

References

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

[2] 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*, pages 22522–22531, 2023.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

[10] 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.

[11] stabilityai. stable-diffusion-3-medium, 2024. URL `https://huggingface.co/stabilityai/stable-diffusion-3-medium`.

[12] 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.

[13] 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)*, pages 10684–10695, June 2022.

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

[15] 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*, pages 2426–2436, 2023.

[16] 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*, pages 1755–1764, 2024.

[17] 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.

[18] 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, pages 6194–6201, 2020.

[19] 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*, pages 270–279, 2017.

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

[21] 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*, pages 1–8, 2011.

[22] 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.

[23] 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)*, pages 4387–4391. IEEE, 2020.

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

[25] 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*, pages 4670–4679, 2021.

[26] 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*, pages 2544–2553, 2022.

[27] 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.

[28] 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.

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

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

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

[32] 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.

[33] Keyu Li, Jin Gao, and Dequan Wang. Aligned agents, biased swarm: Measuring bias amplification in multi-agent systems. *arXiv preprint arXiv:2604.08963*, 2026.

[34] Mohan Jiang, Dayuan Fu, Junhao Shi, Ji Zeng, Weiye Si, Keyu Li, Xuefeng Li, Yang Xiao, Wenjie Li, Dequan Wang, et al. davinci-agency: Unlocking long-horizon agency data-efficiently. *arXiv preprint arXiv:2602.02619*, 2026.

[35] Jin Gao, Juntu Zhao, Keyu Li, and Dequan Wang. Kan-mixer: Kolmogorov-arnold networks for gene expression prediction in plant species. In *European Conference on Computer Vision*, pages 135–150. Springer, 2024.

[36] 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.

[37] 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*, pages 7559–7568, 2024.

[38] 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*, pages 5111–5120, 2024.

[39] 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*, pages 6430–6440, 2024.

[40] 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.

[41] 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.

[42] 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*, 2024.

[43] 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*, pages 22691–22702, 2023.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

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

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

[51] 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*, pages 8748–8763. PMLR, 2021.

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

[53] John Canny. A computational approach to edge detection, 1986.

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

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

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

A Comprehensive Evaluation Metrics

We establish a comprehensive, multi-grained evaluation framework comprising five distinct metrics, each with a stringently defined scope to capture the complex trade-offs between local concept erasure and global image preservation, as outlined in Table 5.

CLIPScore calculates the CLIP similarity between the textual company name query and the isolated logo region extracted from the generated image. A lower score indicates a highly effective semantic decoupling, demonstrating that the generated logo no longer aligns with the targeted brand identity.

LogoScore evaluates the latent CLIP feature similarity between the localized logo regions extracted before and after the unlearning intervention. A lower value implies a profound feature-level disruption, marking a successful localized unlearning process.

LogoSSIM utilizes the Structural Similarity Index Measure (SSIM) to quantify the pixel-level structural alterations within the localized logo bounding box before and after unlearning. A diminished SSIM value confirms that the geometric and textural integrity of the logo has been thoroughly dismantled.

ImageScore computes the global CLIP similarity across the entire image composition before and after the unlearning process. A higher score is imperative here; it serves as a critical counterbalance, signifying that the broad contextual semantics and background elements remain fundamentally undisturbed by the localized unlearning.

ImageSSIM measures the global SSIM for the complete visual scene before and after unlearning. A higher value dictates rigorous structural preservation of the non-target background, which is the hallmark of a non-destructive, highly precise unlearning methodology.

These metrics constitute the rigorous backbone of our evaluation methodology, providing a mathematically structured approach to assessing unlearning performance. Furthermore, a comprehensive visual schema of our evaluation pipeline is depicted in Figure 7 to ensure absolute methodological transparency.

Formally, let I_1, I_2, and I_3 denote the original synthesized images, while their corresponding counterparts post-unlearning are denoted as I_a, I_b, and I_c. Following the application of our open-vocabulary logo detection, the localized bounding box regions yielding the highest confidence are extracted and denoted as L_1, L_2, L_3, and L_a, L_b, L_c respectively. CLIPScore calculates the CLIP cosine similarity between the localized sets $\{L_1, L_2, L_3, L_a, L_b, L_c\}$ and the explicit text query (e.g., "apple logo"). LogoScore computes the pairwise CLIP feature similarity between the pre- and post-unlearning local regions: $(L_1, L_a), (L_2, L_b)$, and (L_3, L_c). LogoSSIM measures the pixel-wise SSIM over the same localized pairs: $(L_1, L_a), (L_2, L_b)$, and (L_3, L_c). To capture global fidelity, ImageScore calculates the holistic CLIP similarity between the full image pairs: $(I_1, I_a), (I_2, I_b)$, and (I_3, I_c). Finally, ImageSSIM computes the global SSIM similarity for $(I_1, I_a), (I_2, I_b)$, and (I_3, I_c).

The algorithmic formulation for this evaluation framework is meticulously detailed in Algorithm 1. Both CLIP and SSIM metrics are initially computed per individual prompt by contrasting the pre- and post-intervention states, followed by rigorous aggregation at both the prompt and company levels to ensure statistical robustness.

B 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.

C Empirical Ranking of Logo Entanglement

Beyond benchmarking unlearning algorithms, a pivotal secondary contribution of LU-500 is its capacity to empirically quantify and rank the deep-seated entanglement—or unlearning difficulty—of specific corporate logos within the model's latent space. We formulate this difficulty ranking across the Fortune 500 companies by analyzing the localized disruption metric (LogoScore) under the adversarial conditions of LUim-500 utilizing our exploratory ProLU.

Given that a minimized LogoScore score denotes optimal semantic disruption, an elevated LogoScore score indicates profound resistance to unlearning (i.e., extreme latent entanglement). Sorted in descending order of unlearning resistance, the five most intractably entangled company logos are *HYUNDAI MOTOR*, *CHINA AEROSPACE SCIENCE & INDUSTRY*, *PANASONIC HOLDINGS*, *RTX*, and *TATA MOTORS*, registering extreme LogoScore retention values of $0.834, 0.8299, 0.8142, 0.8122$, and 0.8051 respectively. Conversely, the five most highly malleable logos, demonstrating the least resistance to semantic erasure, are *ORANGE*, *WELLS FARGO*, *AMAZON.COM*, *TARGET*, and *WORLD KINECT*, yielding disrupted LogoScore scores of $0.0907, 0.1838, 0.2119, 0.2159$, and 0.2206.

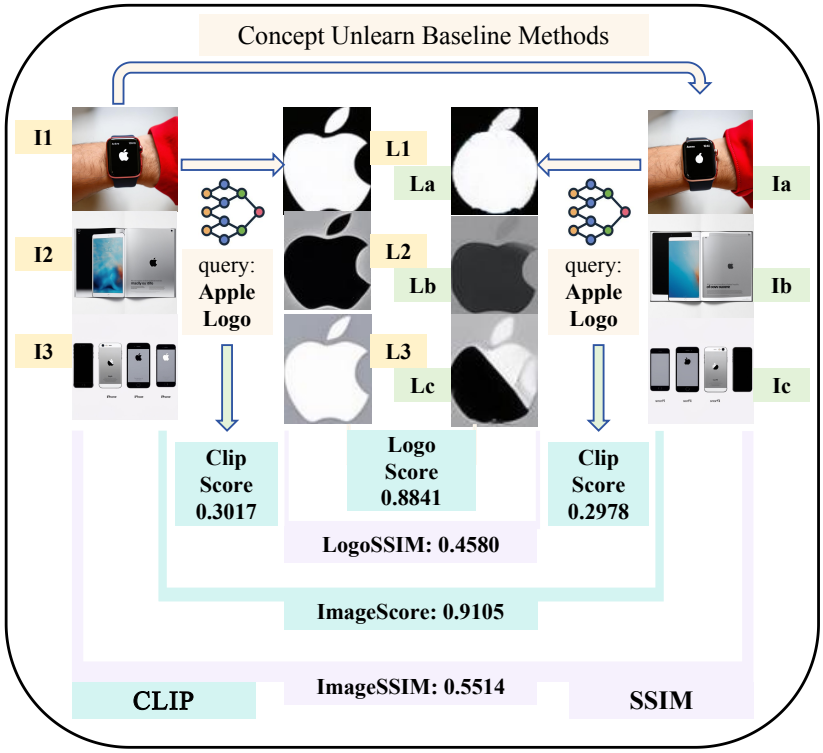

Figure 7: Original images and their unlearned counterparts are analyzed using metrics focused on logos and overall image similarity. CLIPScore calculates the CLIP similarity between detected logos and the text query "apple logo." LogoScore measures the CLIP similarity between logos before and after unlearning. LogoSSIM evaluates the SSIM similarity between logos before and after unlearning. ImageScore assesses the CLIP similarity for the entire image before and after unlearning, while ImageSSIM evaluates the SSIM similarity for the entire image. Light purple represents the SSIM score, while light blue represents the CLIP score.

Algorithm 1 The CLIP and SSIM similarities are first calculated for each individual prompt by comparing the images before and after unlearning. Then, the results are averaged at both the prompt level and the company level.

Require: Datasets D_i, $i = 1, 2$; Unlearning methods M_{il}; Companies C_{ilj}; Prompts P_{iljk}
Ensure: Metrics ME_{il} for evaluating unlearning methods

for each dataset D_i **do**
  for each unlearning method M_{il} **do**
    for each company C_{ilj}, $j = 1$ to 500 **do**
      for each prompt P_{iljk}, $k = 1$ to 10 **do**
        {Generate images before and after unlearning}
        Generate image I_ori_{iljk} using Stable Diffusion before unlearning
        Apply unlearning method M_{il} to generate I_un_{iljk}
        {Metric 1: Text-Logo Alignment (Local View)}
        Use OWLv2 with C_{ilj} logo as text query on I_ori_{iljk} and I_un_{iljk}
        Extract top confidence scores corresponding boxes containing logos as L_ori_{iljk} and L_un_{iljk}
        Extract CLIP logo features $F_L_ori_{iljk}$ and $F_L_un_{iljk}$
        Extract CLIP text features T_{ilj} with text query C_{ilj} logo
        Compute cosine similarity s_0, s_1 between $F_L_ori_{iljk}$, $F_L_un_{iljk}$ and T_{ilj}
        $ME1_{iljk} = s_0$ before unlearn or $= s_1$ after unlearn
        {Metric 2: Logo-Logo Alignment (Local View)}
        Extract CLIP features $F_L_ori_{iljk}$ and $F_L_un_{iljk}$
        Compute cosine similarity $ME2_{iljk}$ between $F_L_ori_{iljk}$ and $F_L_un_{iljk}$
        {Metric 3: Logo-Logo Alignment (Local View)}
        Compute SSIM score $ME3_{iljk}$ between L_ori_{iljk} and L_un_{iljk}
        {Metric 4: Image-Image Alignment (Global View)}
        Extract CLIP features $F_I_ori_{iljk}$ and $F_I_un_{iljk}$
        Compute cosine similarity $ME4_{iljk}$ between $F_I_ori_{iljk}$ and $F_I_un_{iljk}$
        {Metric 5: Image-Image Alignment (Global View)}
        Compute SSIM score $ME5_{iljk}$ between I_ori_{iljk} and I_un_{iljk}
      end for
      {Average metrics over k}
      Compute $ME1_{ilj} = \frac{1}{10}\sum_{k=1}^{10} ME1_{iljk}$
      Compute $ME2_{ilj} = \frac{1}{10}\sum_{k=1}^{10} ME2_{iljk}$
      Compute $ME3_{ilj} = \frac{1}{10}\sum_{k=1}^{10} ME3_{iljk}$
      Compute $ME4_{ilj} = \frac{1}{10}\sum_{k=1}^{10} ME4_{iljk}$
      Compute $ME5_{ilj} = \frac{1}{10}\sum_{k=1}^{10} ME5_{iljk}$
    end for
    {Average metrics over j}
    Compute $ME1_{il} = \frac{1}{500}\sum_{j=1}^{500} ME1_{ilj}$
    Compute $ME2_{il} = \frac{1}{500}\sum_{j=1}^{500} ME2_{ilj}$
    Compute $ME3_{il} = \frac{1}{500}\sum_{j=1}^{500} ME3_{ilj}$
    Compute $ME4_{il} = \frac{1}{500}\sum_{j=1}^{500} ME4_{ilj}$
    Compute $ME5_{il} = \frac{1}{500}\sum_{j=1}^{500} ME5_{ilj}$
  end for
end for
return ME_{il} for all methods M_{il}

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		✓		✓		✓

D Prompt Curation and Multi-Agent Pipeline

To ensure absolute methodological transparency in the construction of LU-500, we disclose the exact meta-prompts utilized to synthesize the dataset. Capitalizing on the advanced reasoning capabilities of state-of-the-art language models, these were generated leveraging OpenAI's GPT-5, as thoroughly documented in Figure 8. Furthermore, we transparently detail the architectural prompt logic underpinning our exploratory multi-agent baseline (ProLU), which orchestrates the collaborative functionalities of the Remover, Reflector, and Checker agents, visually formalized in Figure 9.

Within this semantic intervention pipeline: The *Remover* acts as the primary semantic ablator, engineered to surgically excise any latent textual triggers linked to the target logo while meticulously preserving the surrounding contextual phrasing. The *Reflector* functions as a semantic stabilizer; it critically evaluates the Remover's output against the original user intent, iteratively optimizing the prompt to correct over-edits or hallucinated contexts. The *Checker* provides a deterministic final safeguard, executing a rigorous review to certify the absolute absence of implicit or explicit brand references. Should residual triggers be detected, the prompt is immediately cycled back for aggressive, direct ablation.

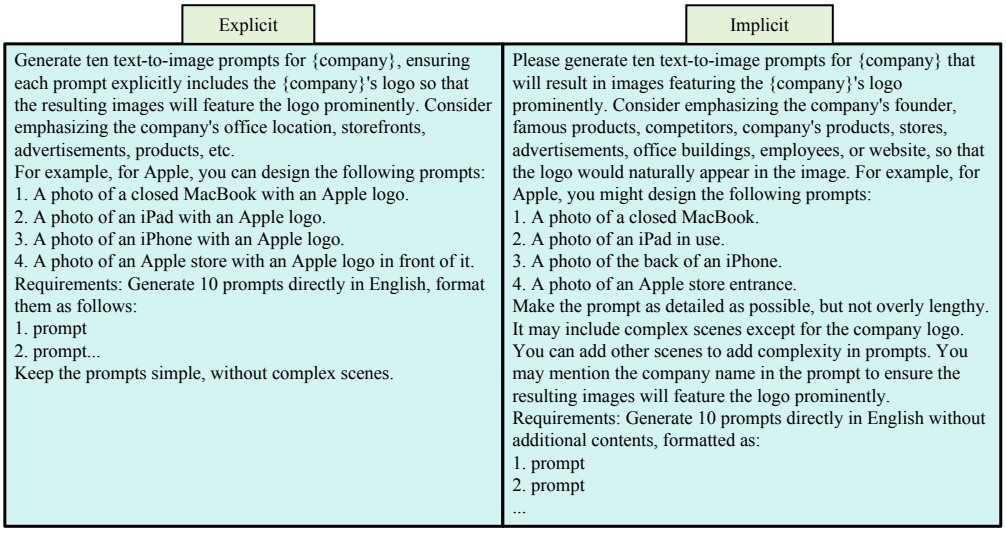

Figure 8: The agent prompts for generating LU-500 was crafted using OpenAI's GPT-4o model.

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: The Remover, Reflector, and Checker prompts in ProLU. The Remover is used to eliminate elements related to the company logo from the original prompt while keeping other parts as consistent as possible. The Reflector evaluates whether the Remover has successfully completed its task and provides further optimized prompts. The Checker performs a final review to ensure that the final prompt does not contain any company logo; if any logo-related elements remain, they are directly removed.

E Extended Qualitative Evaluation

Finally, to provide an exhaustive demonstration of unlearning dynamics across a diverse topological and semantic spectrum, we supply an extended gallery of supplementary visual results. These comprehensive qualitative evaluations are indexed across Figure 10 and Figure 11, serving to further validate the behavioral boundaries of our approach.

To ensure an unbiased and comprehensive representation of the Fortune Global 500 distribution—encompassing both text-heavy and geometrically complex designs—we systematically sampled one representative company for nearly every letter of the alphabet from the LU-500 corpus (excluding 'Y', which contained no valid entries). The evaluated cohort explicitly includes: Apple, Boeing, Coca-Cola, DELL, EXXON MOBIL, FedEx, Goldman-Sachs, HP, Intel, Johnson & Johnson, KIA, L'Oreal, Mercedes-Benz, Nike, Oracle, Pfizer, Qualcomm, Renault, Starbucks, Tesla, Uber, Volvo, Walt Disney, Xiaomi, and Zurich Insurance Group.

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

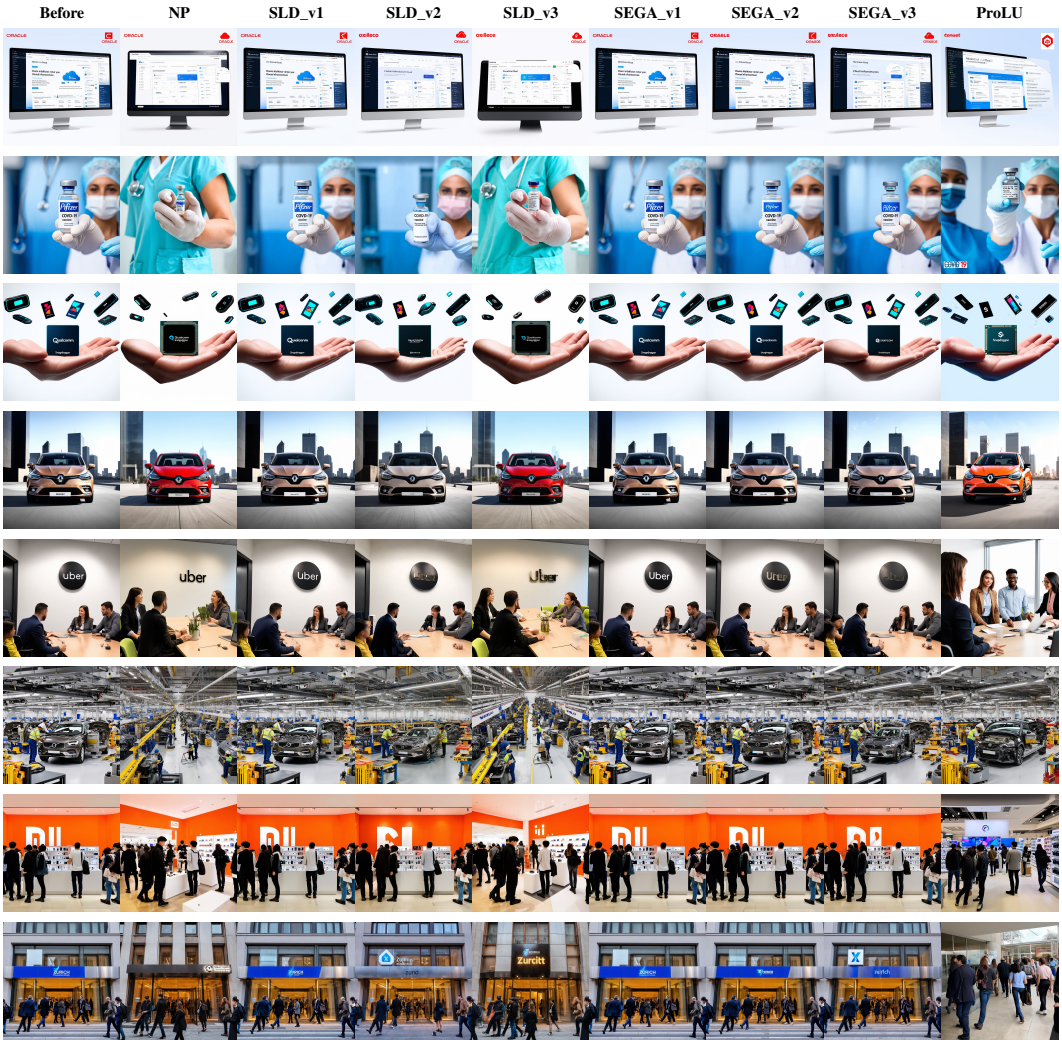

Figure 11: More visual results on Oracle, Pfizer, Qualcomm, Renault, Uber, Volvo, Walt Disney, Xiaomi, and Zurich Insurance Group.