Entangled: A Unified Metric for Evaluating Selective Forgetting in Generative Models

Anonymous ICCV submission

Paper ID 19

Abstract

The widespread deployment of diffusion models in image generation raises growing concerns about privacy-compliant selective forgetting. Existing methods often fail to remove sensitive content without compromising the quality of surrounding regions, largely due to the entangled and high-dimensional nature of generative latent spaces. To address this, we propose Entangled, a unified metric that jointly quantifies forgetting effectiveness and retention consistency by measuring semantic disassociation and structural fidelity between target and background regions. Entangled supports both paired (Entangled-D) and unpaired (Entangled-S) evaluation, enabling unsupervised assessment of unlearning behavior. To facilitate evaluation, we further introduce a training-free framework that simulates forgetting via prompt-driven layered editing, yielding a lightweight benchmark dataset called ForgetMe. It includes both real and synthetic samples across ImageNet, Stanford Dogs, CUB-200-2011, and a generated cat dataset. Experimental results on LoRA-finetuned Stable Diffusion validate the reliability of both Entangled and ForgetMe as scalable tools for assessing selective forgetting. Our work offers a robust and generalizable foundation for privacy-preserving generative modeling. Code and dataset will be released upon acceptance.

1. Introduction

Selective forgetting in generative models refers to the ability to remove specific concepts or visual elements from model outputs while preserving unrelated content [18, 19]. This capability is increasingly important for privacy compliance (e.g., GDPR), error correction, and controlled editing in large-scale models [32, 37]. Unlike in supervised classification, where forgetting is often defined via gradient erasure or fine-tuning [29, 31, 34], generative models such as diffusion models pose unique challenges due to their high-dimensional outputs and entangled latent representations [11].

Prior works on object removal and concept erasure [9, 33] have demonstrated partial success in removing target information. However, current evaluation metrics fail to reflect the dual objectives of forgetting: (1) ensuring the target content is effectively removed, and (2) preserving the integrity of the remaining content. For example, FID [14] measures global distributional shift, CLIP Distance [15] focuses on semantic alignment, and PSNR/SSIM [13] assess low-level fidelity—but none simultaneously quantify deletion success and retention consistency.

To fill this gap, we propose a new evaluation metric, *Entangled*, specifically designed for selective forgetting in generative models. Our goal is to provide a unified, model-agnostic, and interpretable metric that enables principled comparisons of forgetting methods across various tasks and datasets. Our **contributions** are as follows:

- **Entangled Metric.** We introduce *Entangled*, a unified metric that quantifies both forgetting success and background consistency for generative models.
- **Training-Free Benchmark.** We construct a lightweight dataset using layered image editing to simulate controlled forgetting scenarios.
- **Empirical Validation.** We evaluate Entangled on cat and bird categories, demonstrating its effectiveness compared to existing metrics and baselines.

2. Related Work

2.1. Selective Forgetting

Selective forgetting refers to the targeted removal of specific concepts or features from a generative model's outputs, without retraining the entire model [8]. This capability is increasingly important for privacy protection—such as compliance with the GDPR's "right to be forgotten"—as well as for ensuring safety, fairness, and controllability in content generation [1, 2, 20].

Challenges in Diffusion Models. In diffusion models, selective forgetting poses unique challenges. These models generate outputs by denoising from latent representa-

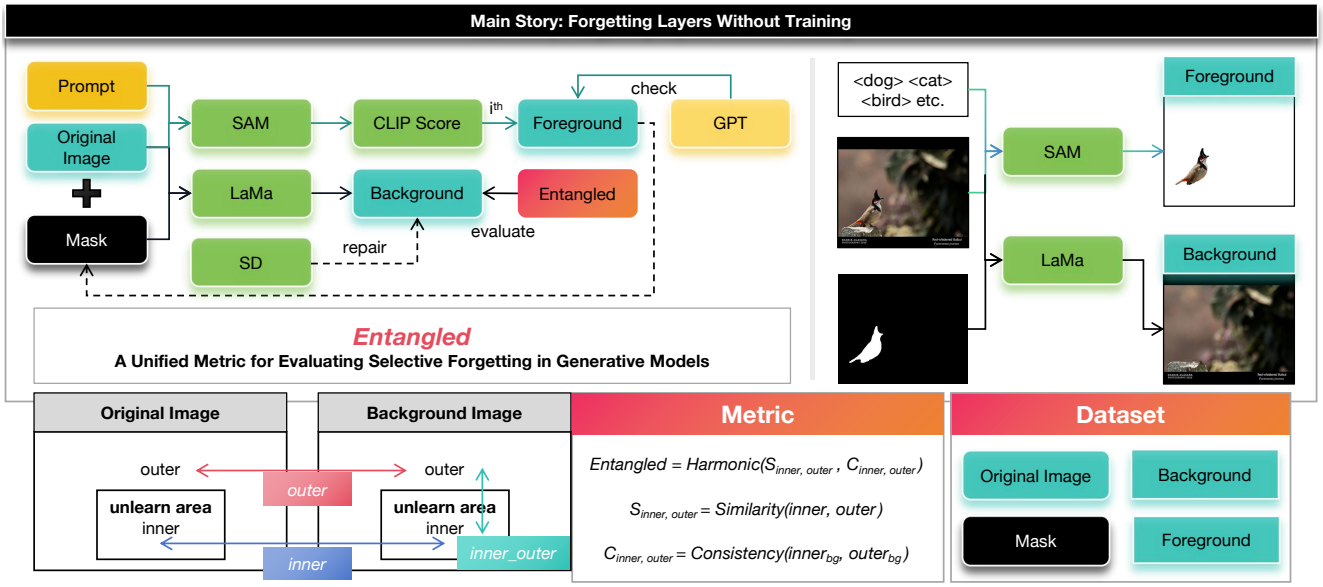

Figure 1. Pipeline for generating evaluation samples used with the *Entangled* metric. **Main Story** applies layered editing to remove target content and reconstruct the background, enabling controlled assessment of selective forgetting.

tions that are high-dimensional, continuous, and often entangled [38]. As a result, removing one concept may inadvertently affect unrelated regions or degrade the coherence of the background [27, 28, 36]. This entanglement makes it difficult to erase specific content without disrupting the overall structure or style.

Concept-Level Forgetting. To address this, recent methods have explored concept-level forgetting by editing the latent space of diffusion models [4, 9, 35], or by fine-tuning select layers using low-rank adaptation (LoRA). While these approaches can suppress targeted concepts, they often introduce unintended changes to global appearance or compromise generation fidelity [23, 39].

Instance-Level Removal. Other works perform object-level removal using spatial masks and text prompts. For example, Inst-Inpaint [3, 33] leverages diffusion-based inpainting to remove specified regions. Although effective in local editing, such methods often produce artifacts like edge discontinuities, texture mismatches, or incomplete removal—especially in complex or high-resolution scenes [10].

Limitations of Existing Methods. While recent methods have made progress in removing concepts or objects from generated content, they often lack precise control over what is forgotten versus retained. More importantly, they are difficult to evaluate in a consistent and interpretable manner. This highlights the need to revisit how forgetting effectiveness is measured—both in terms of content removal and background preservation—which we discuss next.

2.2. Challenges in Evaluation Metrics

Evaluation Metrics. Existing metrics for generative models are not well-suited for evaluating selective forgetting. Frechet Inception Distance (FID) [5, 21] and similar perceptual scores focus on overall image realism but are agnostic to whether specific content has been successfully removed. CLIP-based metrics, such as CLIP Distance and CLIP Maximum Mean Discrepancy (CMMD) [14, 33], evaluate semantic alignment but overlook spatial consistency and the integrity of unedited regions. These limitations are exacerbated by the high dimensionality and prompt sensitivity of diffusion models [25, 30].

Need for Controlled Evaluation. A further challenge lies in the lack of evaluation setups that enable controlled, fine-grained analysis of forgetting behavior. Existing datasets are not tailored for prompt-specific or region-level unlearning tasks, making it difficult to isolate forgetting from other generation variations. This motivates the need for a unified metric that can evaluate forgetting effectiveness and retention fidelity under localized editing — a gap our proposed *Entangled* metric aims to fill.

3. Evaluation Metric - *Entangled*

In removal tasks, our goal is to make the removed regions as distinct as possible, while retaining regions that remain as consistent as possible. This application scenario inspired us to propose a quantitative evaluation metric for selective unlearning, called *Entangled*. This metric assesses the residual association between the foreground and background in the modified images, measuring the thoroughness of the un-

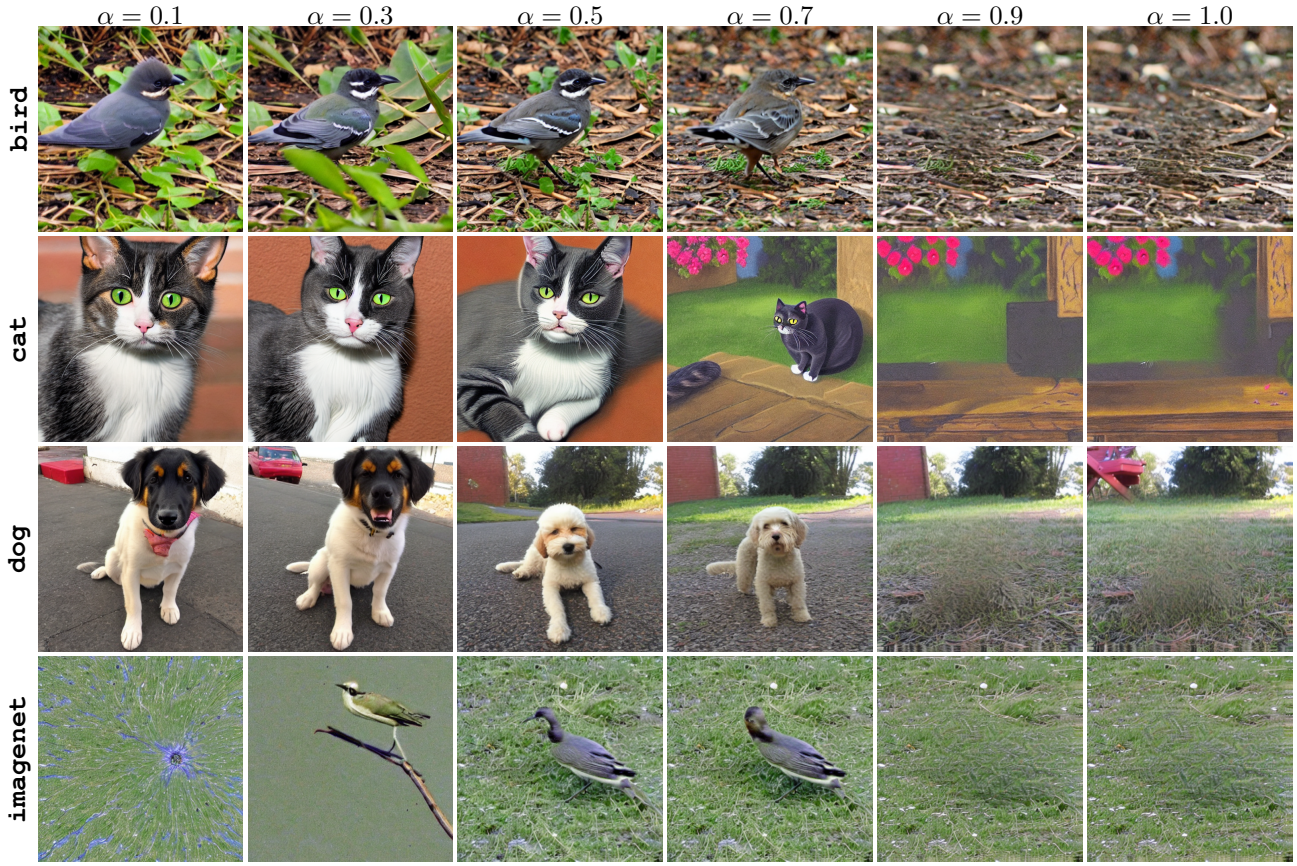

Figure 2. Ablation study on the impact of α in LoRA for selective unlearning across four *ForgetMe* datasets: Bird, Cat, Dog, and ImageNet. $\alpha \in [0.1, 1.0]$ shows that effective unlearning is achieved only when $\alpha \geq 0.9$.

learning process. It combines similarity and consistency metrics to ensure maximum separation between foreground and background information after unlearning.

Definition 3.1 (Entangled Metric)**.** The *Entangled* metric quantifies similarity and consistency between inner and outer regions of original and background images. We define:

$$\begin{aligned} Entangled &= \frac{\alpha + \beta}{\frac{\alpha}{S_{\text{inner, outer}}} + \frac{\beta}{C_{\text{inner, outer}}}} \\ &= \frac{(\alpha + \beta) \cdot S_{\text{inner, outer}} \cdot C_{\text{inner, outer}}}{\alpha \cdot C_{\text{inner, outer}} + \beta \cdot S_{\text{inner, outer}}} \end{aligned} \tag{1}$$

where $S_{\text{inner, outer}}$ is the combined similarity, and $C_{\text{inner, outer}}$ is the consistency metric, with weights $\alpha, \beta \geq 0$ such that: $\alpha + \beta = 1$. Adjusting α and β prioritizes either similarity or consistency; setting $\alpha = 0$ and $\beta = 1$ emphasizes consistency in single-image scenarios. The default $\alpha = \beta = 0.5$ balances both aspects.

Definition 3.2 (Similarity)**.** Let X and Y be pixel matrices for original and background images, normalized to $X, Y \in [0, 1]$. Define regions Ω_{inner} and Ω_{outer} with pixels $X_i, Y_i \in \Omega_{\text{inner}}$ and $X_j, Y_j \in \Omega_{\text{outer}}$.

$$S_{\text{region}} = \sqrt{\frac{1}{N_{\text{region}}} \sum_{k \in \Omega_{\text{region}}} (X_k - Y_k)^2} \tag{2}$$

where $N_{region} \in N_{inner}, N_{outer}$, $k \in i, j$, and N is total pixel counts.

To ensure consistency with a monotonically increasing relationship, regardless of proximity, the outer similarity is adjusted as $1 - S_{\text{outer}}$. The combined similarity metric $S_{\text{inner, outer}}$ is defined as:

$$S_{\text{inner, outer}} = \frac{2 \cdot S_{\text{inner}} \cdot (1 - S_{\text{outer}})}{S_{\text{inner}} + (1 - S_{\text{outer}}) + \epsilon} \tag{3}$$

where $\epsilon = 1 \times 10^{-6}$ prevents division by zero.

Remark 3.3. The combined similarity metric $S_{\text{inner, outer}}$ attains its maximum value of 1, i.e.,

$$S_{\text{inner, outer}} \to 1 \quad \text{as} \quad S_{\text{inner}} \to 1,\ S_{\text{outer}} \to 0.$$

Conversely, $S_{\text{inner, outer}}$ reaches its minimum value of 0, i.e.,

$$S_{\text{inner, outer}} \to 0 \quad \text{as} \quad S_{\text{inner}} \to 0,\ S_{\text{outer}} \to 1.$$

Definition 3.4 (Consistency)**.** The consistency metric $C_{\text{inner, outer}}$ measures alignment of statistical characteristics (mean and variance) between inner and outer regions to maintain coherence. Define:

$$C_{\text{inner, outer}} = M \times V \tag{4}$$

where mean consistency M and variance consistency V are given by:

$$M = \frac{2 \cdot \mu_{\text{inner}} \cdot \mu_{\text{outer}}}{\mu_{\text{inner}}^2 + \mu_{\text{outer}}^2 + \epsilon} \tag{5}$$

$$V = \frac{2 \cdot \sigma_{\text{inner}} \cdot \sigma_{\text{outer}}}{\sigma_{\text{inner}}^2 + \sigma_{\text{outer}}^2 + \epsilon} \tag{6}$$

where:

$$\mu_{\text{region}} = \frac{1}{N_{\text{region}}} \sum_{k \in \Omega_{\text{region}}} X_k \tag{7}$$

$$\sigma_{\text{region}}^2 = \frac{1}{N_{\text{region}}} \sum_{k \in \Omega_{\text{region}}} (X_k - \mu_{\text{region}})^2 \tag{8}$$

where $\epsilon = 1 \times 10^{-6}$ prevents division by zero.

Remark 3.5. The consistency metric $C_{\text{inner, outer}}$ approaches its maximum value of 1, i.e.,

$$C_{\text{inner, outer}} \to 1 \quad \text{as} \quad |\mu_{\text{inner}} - \mu_{\text{outer}}| \to 0, \quad \text{and} \quad |\sigma_{\text{inner}} - \sigma_{\text{outer}}| \to 0.$$

Higher $C_{\text{inner, outer}}$ indicate a closer match in statistical characteristics, promoting visual coherence. Conversely,

$$C_{\text{inner, outer}} \to 0 \quad \text{as} \quad |\mu_{\text{inner}} - \mu_{\text{outer}}| \to +\infty, \quad \text{or} \quad |\sigma_{\text{inner}} - \sigma_{\text{outer}}| \to +\infty,$$

where the metric approaches its minimum value of 0, suggesting greater statistical discrepancy and potentially incomplete background preservation.

4. Framework for Selective Forgetting

We propose a modular, **training-free** framework that simulates selective forgetting in generative models by removing target content while preserving surrounding semantics—without model retraining. Operating directly on diffusion outputs, it leverages the *Entangled* metric for automated supervision. The process follows a two-stage editing pipeline:

- **Foreground Extraction:** Identifies and isolates the target semantic region for forgetting.
- **Background Reconstruction:** Uses generative inpainting models to restore the masked region with high visual coherence.

4.1. Foreground Extract

The goal of foreground extraction is to isolate the specific object or information that needs to be removed from the image. The steps for this process are as follows:

Step 1: Generate segmentation masks. SAM (Segment Anything Model) [17] is used to produce multiple segmentation masks on the image, dividing it into various regions.

Step 2: Calculate CLIP scores for each mask. Using the CLIP model, a similarity score (CLIP score [12]) is computed for each mask, assessing how closely each region corresponds to the target information.

Step 3: Select the highest-scoring mask and crop the foreground region. The mask with the highest CLIP score is selected, and the corresponding region is cropped as an initial foreground area.

Step 4: Validate the foreground using GPT. The cropped foreground image is input to GPT along with a verification prompt.

Q: Is this a [target category]?

A: Yes, this is a [target category].

Solution: The extraction is considered successful.

A: No, this is not a [target category].

Solution: We return to Step 3 and select the next highest-scoring mask for foreground extraction.

Step 5: Final foreground confirmation. The process repeats until GPT successfully verifies the foreground as belonging to the intended category, designating it as the final foreground selection.

4.2. Background Reconstruction

The goal of background reconstruction is to restore the missing region left by the removed foreground, ensuring visual coherence and structural consistency. The specific steps are as follows:

Step 1: Prepare the image for inpainting. The region corresponding to the removed foreground is masked out, creating an incomplete image with a blank area for inpainting.

Step 2: Initial inpainting with LaMa. The LaMa model [24] is applied to the blank region, performing an initial inpainting to generate a plausible background.

Step 3: Evaluate background quality. The *Entangled* metric is used to assess the reconstructed background by measuring its residual association with the removed foreground. If the score is low, indicating poor reconstruction quality, further refinement is required.

Step 4: Refine prompt and reapply inpainting. For backgrounds with low *Entangled* scores, the inpainting prompt is manually refined to enhance the generation process. Stable Diffusion (SD) [22] is then employed for a second inpainting pass to enhance background coherence and realism.

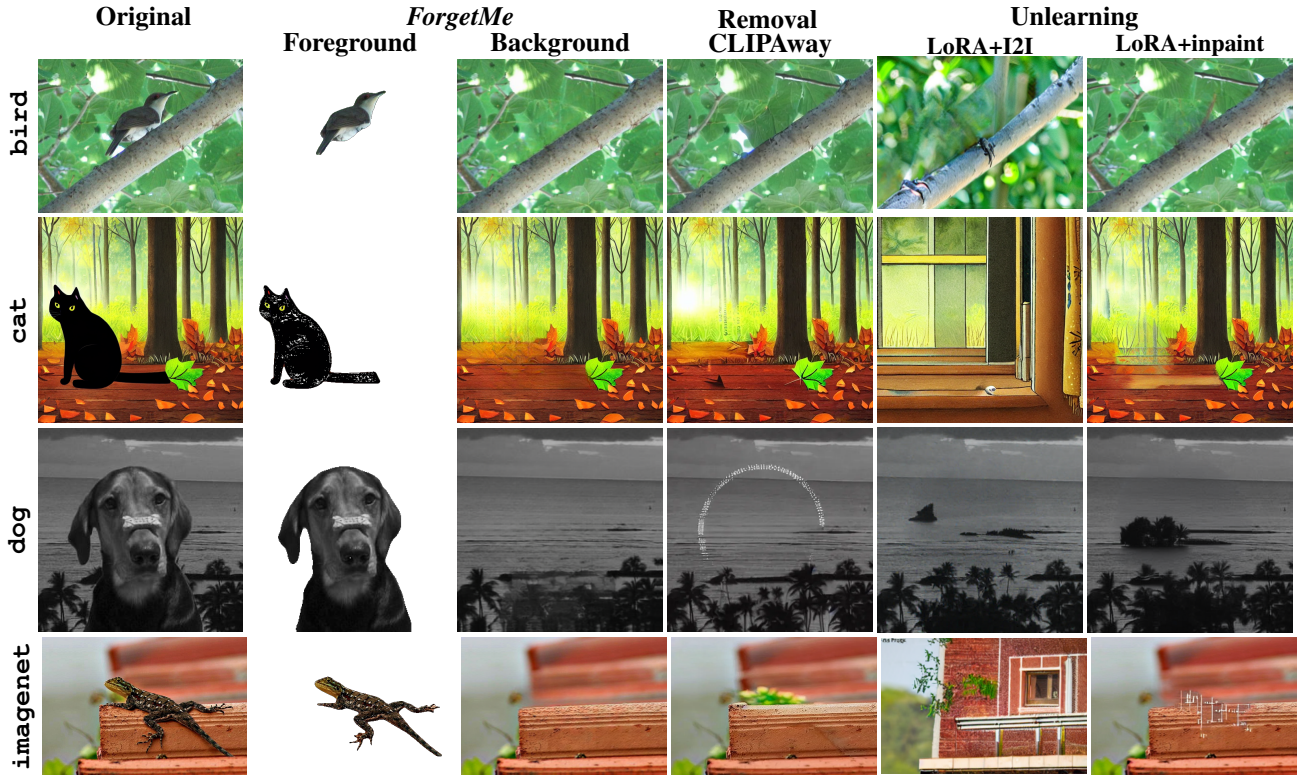

Figure 3. Comparison of model performance on selective unlearning tasks using our proposed dataset (*ForgetMe*), showcasing different categories, including Bird, Cat, Dog, and ImageNet, with both foreground and background layers. Results compare the state-of-the-art Object Removal model CLIPAway [7], and models trained for unlearning using our *ForgetMe* dataset.

5. Experiment

Dataset Description. We evaluate unlearning performance using three public datasets—**ImageNet** [6], **Stanford Dogs** [16], and **CUB-200-2011 (Bird)** [26]—with category-level prompts (e.g., *dog*, *bird*), and a synthetic **Generated-Cats** set of 10,000 images created via Stable Diffusion and GPT-4o prompts. Additionally, we introduce **ForgetMe**, a layered benchmark containing original, foreground, background, and mask for each sample, enabling controlled evaluation with ¿90% foreground removal success.

Evaluation Metrics. We adopt standard metrics to evaluate unlearning performance: **FID** [5] measures image quality via feature statistics; **CMMD** [14] and **CLIP Distance** [33] assess distributional and semantic alignment; **CLIP Accuracy** [7] evaluates unlearning success via zero-shot classification (Top-1/3/5); ***Entangled*** captures both forgetting success and retention consistency through region-level analysis.

Comparison. We validated the *Entangled* metric and analyzed selective unlearning across four datasets. Our framework, incorporating both Foreground and Background components (see Figure 3), outperforms SOTA method CLIPAway [7] in flexibility, requiring no additional training. As models improve, our framework is expected to yield even better results. Baselines LoRA+I2I and LoRA+inpaint showed reconstruction quality comparable to SOTA, confirming the dataset's relevance for selective unlearning.

Table 1 presents a quantitative evaluation of each model across various metrics, including our proposed *Entangled-D* (paired images) and *Entangled-S* (unpaired images) metrics, which assess image fidelity and unlearning success. Our framework achieved the best performance on nearly all metrics, with the lowest FID (58.8808) and CMMD (0.8000), and the highest *Entangled-D* (0.8680) and *Entangled-S* (0.8119), highlighting its superior unlearning efficacy and background consistency.

In particular, our framework achieved CLIP@1, CLIP@3, and CLIP@5 accuracies of 0.8096, 0.7739, and 0.6852, significantly outperforming CLIPAway. Its CLIP Distance reached 0.9456, about 13.6% higher than CLIPAway, reflecting enhanced semantic consistency and object removal. For unlearning across contexts, Entangled-D and Entangled-S reliably assess paired and unpaired images, with our framework excelling in both, demonstrating its versatility and robustness.

The strong performance of LoRA+I2I and LoRA+inpaint validates the effectiveness of our framework

Table 1. Quantitative comparison of selective forgetting performance across models. The proposed *Entangled* metric includes two variants: *Entangled-D* for paired-image evaluation and *Entangled-S* for unpaired-image evaluation. Higher values indicate better forgetting fidelity. **Bold** and underlined entries denote the best and second-best scores, respectively.

Models	FID↓	CMMD↓	CLIP Dist.↑	CLIP@1↑	CLIP@3↑	CLIP@5↑	Entangled-D↑	Entangled-S↑
Ours	**58.8808**	**0.8000**	**0.9456**	**0.8096**	**0.7739**	**0.6852**	**0.8680**	**0.8119**
CLIPAway	68.4728	0.8273	0.8324	0.7493	0.7027	0.6658	0.7920	0.7266
LoRA+inpaint	72.9584	0.8231	0.8211	0.7355	0.6791	0.5842	0.6997	0.6350
LoRA+I2I	101.8833	0.9340	0.7808	0.7194	0.7080	0.5802	0.5603	0.5752

in unlearning tasks and highlights the *Entangled* metric's adaptability and reliability for evaluating unlearning in both paired and unpaired image contexts.

Ablation Study. We investigated the impact of the LoRA scaling factor α on unlearning performance by adjusting its value from 0.1 to 1.0. With a fixed random seed for reproducibility, we observed that unlearning effectiveness improved as α increased. In particular, values above 0.9 showed significant forgetting effects while preserving structural consistency (see Figure 2). This trend was consistent across datasets, indicating that a higher α enhances semantic separation between foreground and background. Our results suggest that setting α close to 1.0 is favorable for rigorous and reliable selective unlearning.

6. Conclusion

We propose *Entangled*, a unified evaluation metric for selective unlearning in generative models, designed to jointly assess forgetting effectiveness and retention stability in both paired and unpaired settings. To facilitate controlled benchmarking, we design a training-free, layered editing framework that constructs diverse unlearning scenarios without model retraining. While a lightweight benchmark dataset is provided to support analysis, our primary contribution lies in the *Entangled* metric, which establishes a scalable and model-agnostic foundation for evaluating privacy-preserving unlearning in generative AI. We hope *Entangled* serves as a foundation for future research in safe and interpretable generative modeling.

References

[1] Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review*, 58(3):90, 2025. 1

[2] Kongyang Chen, Zixin Wang, Bing Mi, Waixi Liu, Shaowei Wang, Xiaojun Ren, and Jiaxing Shen. Machine unlearning in large language models. *arXiv preprint arXiv:2404.16841*, 2024. 1

[3] Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[4] Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. *arXiv preprint arXiv:2409.11219*, 2024. 2

[5] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020. 2, 5

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Yigit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models. *arXiv preprint arXiv:2406.09368*, 2024. 5

[8] Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *European Conference on Computer Vision*, pages 278–297. Springer, 2024. 1

[9] 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. 1, 2

[10] 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. 2

[11] Hanlin Gu, WinKent Ong, Chee Seng Chan, and Lixin Fan. Ferrari: federated feature unlearning via optimizing feature sensitivity. *Advances in Neural Information Processing Systems*, 37:24150–24180, 2025. 1

[12] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4

[13] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 1

[14] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 9307–9315, 2024. 1, 2, 5

[15] Rena Kass, Grace Kumar, V Suzanne Klimberg, Lawrence Kass, Ronda Henry-Tillman, Anita Johnson, Maureen Colvert, Sarah Lane, David Harshfield, Soheila Korourian, et al. Clip migration in stereotactic biopsy. *The American journal of surgery*, 184(4):325–331, 2002. 1

[16] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, 2011. 5

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4

[18] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024. 1

[19] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025. 1

[20] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*, 2024. 1

[21] Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, Xingliang Yuan, and Xiaoning Liu. A survey on federated unlearning: Challenges, methods, and future directions. *ACM Computing Surveys*, 57(1):1–38, 2024. 2

[22] 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*, pages 10684–10695, 2022. 4

[23] Vinith M Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. *arXiv preprint arXiv:2410.08074*, 2024. 2

[24] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 4

[25] Bozhong Tian, Xiaozhuan Liang, Siyuan Cheng, Qingbin Liu, Mengru Wang, Dianbo Sui, Xi Chen, Huajun Chen, and Ningyu Zhang. To forget or not? towards practical knowledge unlearning for large language models. *arXiv preprint arXiv:2407.01920*, 2024. 2

[26] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[27] Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[28] YiMin Wen, Xiang Liu, and Hang Yu. Adaptive tree-like neural network: Overcoming catastrophic forgetting to classify streaming data with concept drifts. *Knowledge-Based Systems*, 293:111636, 2024. 2

[29] Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. Evaluating deep unlearning in large language models. *arXiv preprint arXiv:2410.15153*, 2024. 1

[30] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 2

[31] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023. 1

[32] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024. 1

[33] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 1, 2, 5

[34] Zhenyu Yu and Chee Seng Chan. Yuan: Yielding unblemished aesthetics through a unified network for visual imperfections removal in generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1

[35] Zhenyu Yu, Mohd Yamani Idna Idris, and Pei Wang. When cloud removal meets diffusion model in remote sensing. *arXiv preprint arXiv:2504.14785*, 2025. 2

[36] 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. 2

[37] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024. 1

[38] Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37: 12293–12333, 2024. 2

[39] Jingyu Zhu, Ruiqi Zhang, Licong Lin, and Song Mei. Choose your anchor wisely: Effective unlearning diffusion models via concept reconditioning. In *Neurips Safe Generative AI Workshop 2024*. 2