My Art My Choice: Adversarial Protection Against Image Generation

İlke Demir
Cauth AI
ilke@cauth.ai

Anthony Rhodes
Intel Labs
anthony.rhodes@intel.com

Ram Bhagat
Binghamton University
rbhagat2@binghamton.edu

Nese Alyuz Civitci
Intel Labs
nesealyuz@gmail.com

Sinem Aslan
Intel Labs
sinem.aslan@intel.com

Umur Aybars Çiftçi
Binghamton University
uciftci@binghamton.edu

Abstract

Generative AI enables realistic content production via publicly available interfaces, with diffusion models changing the creator economy through high-quality, low-cost content generation. However, artists are resisting unruly AI since their artwork is leveraged by such models without consent. My Art My Choice (MAMC) learns to generate adversarial "protected" versions of images that "break" diffusion models, with distortion levels controlled by the artist to balance fidelity vs. protection. We experiment across multiple datasets and image-to-image tasks, evaluating in visual, noise, structure, pixel, and generative spaces. A user study with 102 artists shows 98% willingness to use MAMC. Overall, MAMC offers crucial protection for preserving ownership against AI-generated content in a human-centric way.

1 Introduction

Generative modeling applications have expanded from shapes [23, 8, 9] to widespread creative use through deep generative models [2, 12]. These models democratize visual content creation but are trained on internet data without ownership consideration, mimicking specific content, style, and structure of samples. Artists resist unruly AI use [4, 13, 10] as: (1) generative AI creates derivatives without liabilities, (2) models train on their data without permission, and (3) there is no compensation.

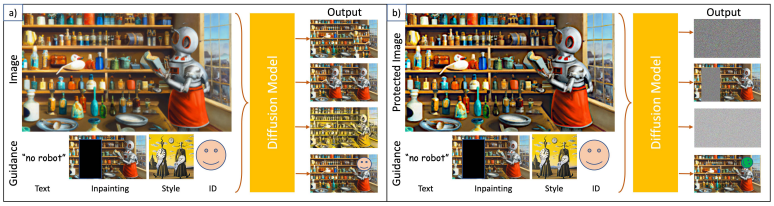


Figure 1: **Motivation.** Given diffusion-guided image generation tasks (a), MAMC aims to confuse DMs for vastly degrading their output (b).

As regulations are immature, we propose "My Art My Choice" - an interim AI tool enabling artists to seal their material with adversarial protection. When imperceptibly altered versions are fed to diffusion models, we aim for degraded output quality (Fig. 1). We model this as a black-box adversarial attack on diffusion models, optimizing for perceptual resemblance between input and protected images, and structural/generative degradation of diffusion outputs. Moreover, MAMC puts artists in control through tunable variables that balance fidelity and robustness. As opposed to traditional adversarial perturbations introducing additive/subtractive noise, we learn adversarial transformations within a larger manifold. Unlike previous work [20, 21], we: (1) do not utilize target

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

styles or concepts but render the output completely distorted, (2) let artists set distortion amounts, and (3) define multi-objective training for higher quality and more atrophy.

Results (Fig. 3) show average 28.5 PSNR and 0.87 SSIM for input-protected image pairs (higher is better), and 17.33 PSNR and 0.40 SSIM for diffusion output comparison (lower is better), demonstrating effective protection. We perform comparisons, ablations with different losses, hyperparameters, user balance parameter levels, applications, and user evaluations.

2 Related Work

Guided Content Generation: Diffusion Models (DMs) [2] surpass GANs [12] in quality and control over generated content. DMs enable personalized generation [16], deepfakes [17], multi-person images [28], object placement [29], image editing [30], and video synthesis [3] - all requiring additional image guidance that MAMC aims to protect.

Adversarial Protection: Most DMs train on large internet datasets without ownership monitoring, enabling replication of content, style, and structure [24, 32, 30]. Emerging research addresses this through machine unlearning [26], style confusion [20], model-specific disabling [27], noise injection [19], compartmentalizing [11], and biometric manipulation [6, 25, 7]. Similarly, MAMC provides adversarial protection by: (1) learning imperceptible adversarial twins, (2) using robust multi-objective training, and (3) providing external controllable balancing, without limitation to specific tasks, models, or domains.

3 My Art My Choice

We assume content owners want to share images online without generative model exploitation. They introduce imperceptible adversarial changes before release, with control over distortion levels. The adversary accesses images without awareness of protection mechanisms. Both parties can query pre-trained DMs with reasonable compute resources.

Given an image I, MAMC learns $G(I) = I + \delta = I'$ where δ is the learned perturbation to attack a black-box DM M. This attack should create maximally dissimilar $M(I)$ and $M(I')$. Thus, adversarial protection optimization becomes

$$\max_{\delta_I} ||M(I + \delta_I) - M(I)||, \;\; s.t. |\delta_I| < \phi_I + \epsilon \tag{1}$$

where ϕ is the balance factor and ϵ represents a small neighborhood. Intuitively, we strive to push the DM output as far as possible from actual output, enabling generalization across tasks and models. We employ a UNet architecture [15] with blocks containing two convolutional layers plus up/downsampling, with encoder/decoder concatenations (Fig. 2). We use a standard frozen pre-trained DM [14] to infer input-output relations, later using the same model in transfer attacks to break other DMs. In order to train MAMC, we formulate a multi-objective function that combines additional losses to satisfy our initial assumptions.

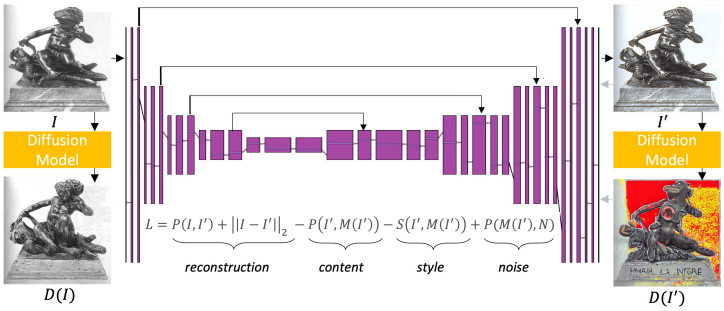


Figure 2: **System Overview.** Protected image I' (top right) is learned from input image I (top left) to break the DM output $M(I)$ (bottom left) as $M(I')$ (bottom right). Generator architecture and loss formulation is also depicted.

Reconstruction Loss: As artists expect minimal changes, we introduce a reconstruction term L_R, keeping I and I' perceptually similar, using LPIPS [31] ($\mathcal{P}$). We also add a pixel-wise ℓ_2 norm.

$$L_R = \alpha_p \mathcal{P}(I, I') + \alpha_r ||I - I'||_2^2 \tag{2}$$

Content Loss: For inpainting and personalization tasks, content should *not* be preserved. We introduce a content loss where $M(I')$ is not perceptually similar to I'.

$$L_C = -\alpha_c \mathcal{P}(I', M(I')) \tag{3}$$

Style Loss: In order to prohibit style transfer and reconstruction tasks, we define style loss as distances between Gram Matrices Ω of I' and $M(I')$, over activations j.

$$L_S = -\alpha_s \frac{1}{|j|} \sum_j ||\Omega_j(I') - \Omega_j(M(I'))|| \tag{4}$$

Noise Loss: Tto really confuse M, we add a noise loss pushing $M(I')$ towards Gaussian noise $\mathcal{N}$.

$$L_N = \alpha_n \mathcal{P}(M(I'), \mathcal{N}) \tag{5}$$

Our final combined loss is constructed as $L = L_R + L_C + L_S + L_N$ where α_* are the loss weights. We provide interactive control with predefined α_* sets, for balancing distortion vs. protection.

4 Results

We evaluate MAMC on three datasets: Wiki Art [18] (1K/5K subsets), Historic Art [1] (1K/5K) and Art 201 [18] (200 images). We select these as representative datasets, with diverse contents, styles, artists, and domains. In all experiments, we normalize images to [0,1] and resize to 512x512. We set 0.001 LR, use Adam, train on GTX 1080 TI, for approximately 1 and 6 hours for 1k and 5k subsets.

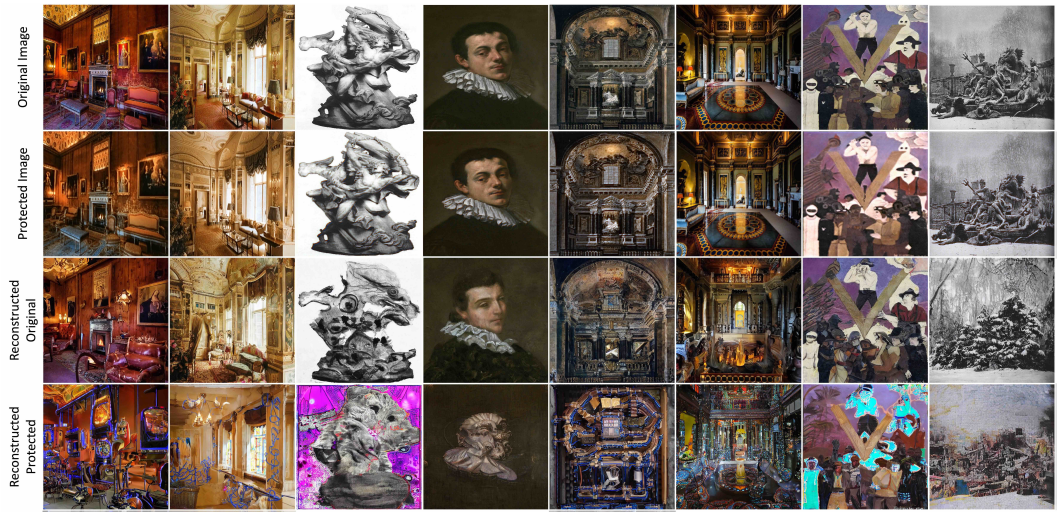


Figure 3: **MAMC Samples.** We enable artists to protect their content (first row) by learning to generate protected versions (second row). Diffusion models mimic content, style, and structure of the art (third row), however, protected images break these models by decreasing output quality (last row).

We want to validate that (1) I and I' are similar and (2) $M(I')$ has low quality. We visualize the success in Fig. 3 and document quantitative evaluations in Tab. 1 in terms of the average PSNR, RMSE, SSIM, and FID for (1) and (2); over three datasets. For (1), we achieve high scores indicating successful preservation. For (2), significantly worse scores confirm diffusion outputs act as adversarial samples with no representative power. MAMC protects any image, independent of training sets. To support this generalization claim, we perform cross-dataset evaluations in Tab. 2.

Comparison: Comparing against image cloaking approaches [22, 5, 19, 20] in Fig. 4, MAMC causes significantly degraded DM outputs. Other methods create plausible outputs that bad actors might still distribute as artists' style or intended concepst for the untrained audiences, while MAMC protected outputs are obviously unusable, in the style of nobody.

Style Protection: Evaluating on single-artist datasets (Edouard Manet, Francesco Albani), MAMC prevents style replication that artists claim steals their identity, as in Fig. 5 along with dataset scores.

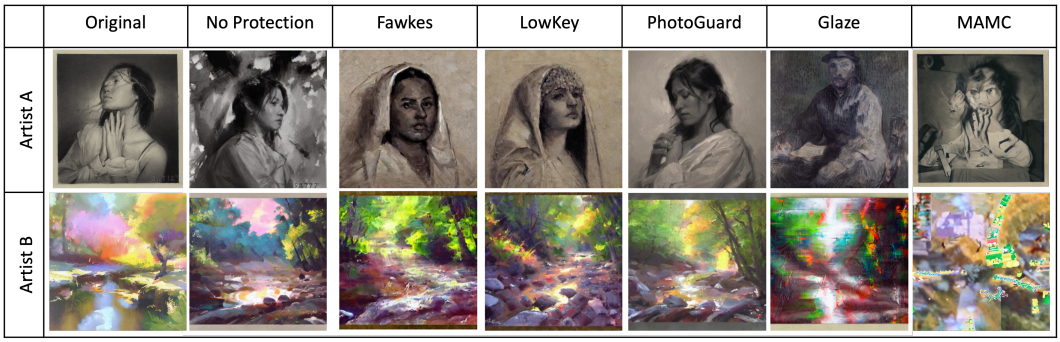


Figure 4: **Comparison.** MAMC not only prevents style mimicking, but also renders outputs useless.

Inpainting: We evaluate MAMC on three scenarios: (1) testing the current pretrained model for inpainting, (2) training and testing our generator for inpainting, and (3) testing this new model on the old task of reconstruction. In Fig. 6, each box contains $I, I', M(I)$ and $M(I')$. Below each, we document same metrics as before. MAMC disables inpainting $M(I')$ almost as effectively as content replication, with additional protection when specifically trained for this task.

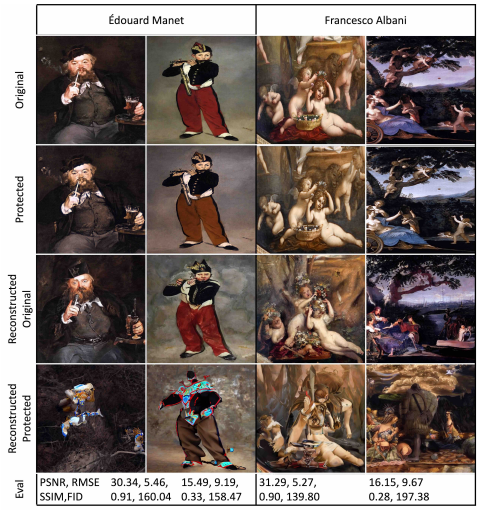


Figure 5: Figure 5: **Style Protection.** DMs fail to replicate style from I' as shown for two artists. Quantitative metrics are per dataset.

		PSNR	RMSE	SSIM	FID
Wiki Art	(1)	25.98↑	7.90↓	0.87↑	123.5↑
	(2)	14.97↓	9.66↑	0.26↓	158.0↓
Hist. Art	(1)	28.15↑	6.36↓	0.88↑	92.83↑
	(2)	16.24↓	9.42↑	0.32↓	163.8↓
Art 201	(1)	24.83↑	7.79↓	0.80↑	209.0↑
	(2)	15.73↓	9.68↑	0.29↓	241.4↓
Face For.	(1)	35.06↑	3.82↓	0.95↑	75.15↑
	(2)	22.40↓	8.81↑	0.73↓	106.9↓

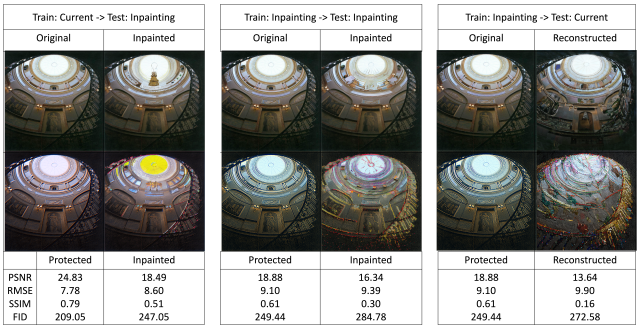
Table 1: Table 1: **Quantitative Evaluation.** Similarities of I and I' (1) and $M(I)$ and $M(I')$ (2).

Train		Wiki Art		His. Art		Art 201	
Test		(1)↑↓↑	(2)↓↑↓	(1)	(2)	(1)	(2)
W. Art	P	25.0	14.9	28.2	16.7	27.5	15.1
	R	7.90	9.66	6.41	9.48	7.17	9.66
	S	0.87	0.26	0.86	0.33	0.89	0.27
His. Art	P	28.2	16.7	28.1	16.2	28.2	16.7
	R	6.41	9.48	6.36	9.42	6.41	9.48
	S	0.86	0.33	0.88	0.32	0.86	0.33
Art 201	P	27.5	15.1	28.2	16.7	24.8	15.7
	R	7.17	9.66	6.41	9.48	7.79	9.68
	S	0.89	0.27	0.86	0.35	0.80	0.29

Table 2: Table 2: **Cross-Dataset Evaluations.** Same metrics with different train/attack datasets.

Ablations: Noise loss restricts representative power; pixel-wise regularization preserves structure and content; style loss provides visual improvements though quantitatively less impactful. Higher diffusion strength incorporates less guidance, requiring less protection as imitation pressure decreases.

Artist Study: We conduct a study with 102 artists understanding perception of protection-caused differences. Most were concerned about generative AI in art (75%). Study involved 20 tasks on 49 unique image pairs, evaluating on 5-point Likert scale. Experiments compared same images (reliability), I vs. $M(I)$ (baseline), I vs. I' (protection similarity), and I vs $M(I')$ (protection effectiveness). Reliability checks showed 93% perceived no difference. Comparing I and $M(I)$, 78% perceived limited difference. However, comparing I and $M(I')$, 74% perceived clear differences, changing to 88%/58% with high/low protection. Protection consistently increased perceived differences across style, content, structure, semantics, emotion, and texture aspects (Fig. 7). For I vs. I', 92% perceived no or very limited difference. With high fidelity, 66% saw no difference. Overall, 56% of participants would use I' over I online, increasing to 96% for high-fidelity protection.



	Protected	Inpainted
PSNR	24.83	18.49
RMSE	7.78	8.60
SSIM	0.79	0.51
FID	209.05	247.05

	Protected	Inpainted
	18.88	16.34
	9.10	9.39
	0.61	0.30
	249.44	284.78

	Protected	Reconstructed
	18.88	13.64
	9.10	9.90
	0.61	0.16
	249.44	272.58

Figure 6: **Inpainting Protection.** Previous model tested for inpainting (left), retrained for inpainting (mid), tested back for reconstruction (right). Full dataset scores below.

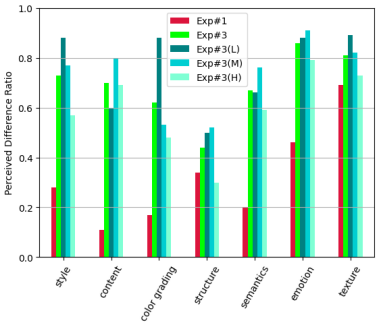


Figure 7: **User Study.** I vs. $M(I)$ for Exp#1 and I' vs. $M(I')$ for Exp#3, per fidelity. Higher = more atrophy.

5 Conclusion

We present "My Art My Choice," an adversarial protection model preventing image exploitation by DMs. With undeniable need for copyright protection, MAMC interrupts personalization, style transfer, inpainting, and guided generation, as evaluated across four datasets with ablations, comparisons, generalizations, and artist perspective. As generative AI emerge, proactive MAMC-based protection will prove valuable. Future work will apply MAMC to other modalities for provenance protection and invest in red teaming for stronger adversarial protection against unseen adversaries.

References

[1] user ansonnn, K.: Historic art dataset. https://www.kaggle.com/datasets/ansonnnnn/historic-art (2023), accessed: 2023-08-10

[2] AUTOMATIC1111: Stable diffusion web ui. `https://github.com/AUTOMATIC1111/stable-diffusion-webui` (2022)

[3] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22563–22575 (June 2023)

[4] Brittain, B.: Getty images lawsuit says stability ai misused photos to train ai. `https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/` (2023), accessed: 2023-08-10

[5] Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., Goldstein, T.: Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. arXiv preprint arXiv:2101.07922 (2021)

[6] Ciftci, U.A., Yuksek, G., Demir, I.: My face my choice: Privacy enhancing deepfakes for social media anonymization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1369–1379 (2023)

[7] Ciftci, U.A., Tanriverdi, A.K., Demir, I.: My body my choice: Human-centric full-body anonymization. arXiv preprint arXiv:2406.09553 (2024), `https://arxiv.org/abs/2406.09553`

[8] Demir, I., Aliaga, D.G., Benes, B.: Proceduralization for editing 3d architectural models. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 194–202. IEEE (2016)

[9] Ebert, D.S., Rohrer, R.M., Shaw, C.D., Panda, P., Kukla, J.M., Roberts, D.A.: Procedural shape generation for multi-dimensional data visualization. Computers & Graphics **24**(3), 375–384 (2000)

[10] Edwards, B.: Artists stage mass protest against ai-generated artwork on artstation. https://arstechnica.com/information-technology/2022/12/artstation-artists-stage-mass-protest-against-ai-generated-artwork/ (2023), accessed: 2023-08-10

[11] Golatkar, A., Achille, A., Swaminathan, A., Soatto, S.: Training data protection with compositional diffusion models. arXiv preprint arXiv:2308.01937 (2023)

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

[13] Lee, T.B.: Stable diffusion copyright lawsuits could be a legal earthquake for ai. https://arstechnica.com/tech-policy/2023/04/stable-diffusion-copyright-lawsuits-could-be-a-legal-earthquake-for-ai/ (2023), accessed: 2023-08-10

[14] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)

[15] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)

[16] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)

[17] Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. arXiv preprint arXiv:2307.06949 (2023)

[18] Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855 (2015)

[19] Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing. arXiv preprint arXiv:2302.06588 (2023)

[20] Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models. arXiv preprint arXiv:2302.04222 (2023)

[21] Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., Zhao, B.Y.: Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In: 2024 IEEE Symposium on Security and Privacy (SP). pp. 807–825. IEEE (2024)

[22] Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., Zhao, B.Y.: Fawkes: Protecting privacy against unauthorized deep learning models. In: 29th USENIX security symposium (USENIX Security 20). pp. 1589–1604 (2020)

[23] Snyder, J.M.: Chapter 6 - applying interval methods to geometric modeling. In: Snyder, J.M. (ed.) Generative Modeling for Computer Graphics and CAD, pp. 163–217. Academic Press (1992). https://doi.org/https://doi.org/10.1016/B978-0-12-654040-6.50012-3, https://www.sciencedirect.com/science/article/pii/B9780126540406500123

[24] Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6048–6058 (2023)

[25] Sunderhaft, A., Bhagat, R., Birchwood, J., Heller, E., Demir, u., Çiftçi, U.A.: Black box adversarial face transformation network. In: Advances in Visual Computing: 19th International Symposium, ISVC 2024, Lake Tahoe, NV, USA, October 21–23, 2024, Proceedings, Part I. p. 330–341. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-77392-1_25, https://doi.org/10.1007/978-3-031-77392-1_25

[26] Tarun, A.K., Chundawat, V.S., Mandal, M., Kankanhalli, M.: Fast yet effective machine unlearning. IEEE Transactions on Neural Networks and Learning Systems pp. 1–10 (2023). https://doi.org/10.1109/TNNLS.2023.3266233

[27] Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. arXiv preprint arXiv:2303.15433 (2023)

[28] Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. arXiv preprint arXiv:2305.10431 (2023)

[29] Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18381–18391 (June 2023)

[30] Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)

[31] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

[32] Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10156 (2023)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are quantitatively and qualitatively supported in results section, also with an additional artist study.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Due to limited space, we omitted discussion of limitations. One limitation we discussed and experimented on are about transfer attacks, which drops protection amount but still effective up to some degree. Another aspect, which can be considered both a limitation or a feature, is the aesthetically diverse results when the fidelity vs. protection knob is tuned.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theoretical proofs are included in 4 pages.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All loss terms, architectural details, and datasets for training/testing are shared; in addition to training configurations.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: While we are not sharing our code, we believe we have shared all the implementation details within the paper to make it reproducible if one wishes to do so. Sharing our code would ease attacking our protection algorithm, thus, we value having a closed source model as another layer of security.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, see the first paragraph of results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We omitted statistical deviations due to space concerns but all full-, single-, or cross-dataset experiments allow us to compute std dev in addition to the reported averages.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the first paragraph of Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we followed NeurIPS code of conduct.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Due to the nature of the topic we have provided the broader impact and positive effects of our work distributed in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The main safeguard we employed is exposing the control for the amount of distortion in artists' content, as one size of protection does not fit all. Since there is no possible misuse scenario, we do not expand with more safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We followed the use rules for licenses of datasets and libraries, citing them as per their instructions. We also obtained necessary permissions for research purposes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: -

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: For our artist study, we could not include all details due to space limitation. They were properly compensated for their work and necessary institutional approvals, workflows, and principles were followed during the study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Again, due to space concerns we have not added these details but separate policies for privacy, data collection, and data retention are followed. Institutional approvals for conducting this study, for working with the third party agency, and for obtaining the anonymized data were properly obtained.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.