Appendix

The appendix is organized as follows:

- In Sec. A1, we provide additional quantitative results. Additionally, we include results of IMMA against the adaptation on multiple concepts in Fig. A10 and Fig. A11.
- In Sec. A2, we provide additional qualitative results. We have also included interactive results (in HTML) along with the supplemental materials.
- In Sec. A3, we document the details of our conducted user study.
- In Sec. A4, we provide additional experimental details, *e.g.*, model architecture, hyperparameters, and description of baseline. We have also attached the code in the supplementary materials and will release the code.

A1 Additional quantitative results

A1.1 Comparison with data poisoning method

We provide quantitative results of MIST [27], a data poisoning method for defending against adaptation. We show the CLIP values after Textual Inversion adaptation in Fig. A1. We observe a gap between the two lines in *purse* and *glasses* which indicates MIST can prevent the model from learning personalized concepts in some datasets. However, compared with IMMA, MIST is less robust and fails in other datasets, *e.g.*, *car*.

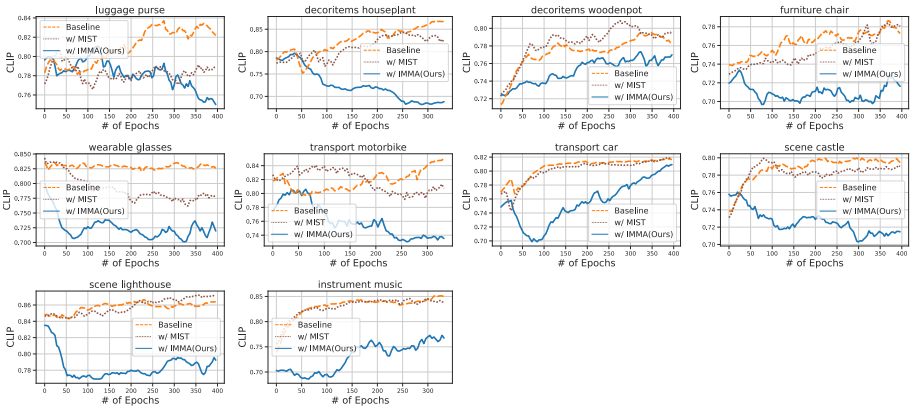

Fig. A1: CLIP versus reference images after Textual Inversion Adaptation.

A1.2 Additional quantitative results of IMMA

Results on immunizing erased model against re-learning. In Fig. A2, we show the metric values *vs*. the number of LoRA adaptation epochs on eight artistic

style datasets. We can observe a gap between the two lines, which indicates that implementing IMMA immunizes the model from re-learning the target artistic style. In Fig. A3, we show the metric values *vs*. the number of LoRA adaptation epochs on ten objects from ImageNet. We can also see a consistent gap between the two lines across all datasets.

Results on immunizing against personalized content. For personalization adaptation, we show more quantitative results on ten datasets from Kumari et al. [24]. The metric values of Textual Inversion, Dreambooth, and Dreambooth LoRA are shown in Fig. A4, Fig. A5, and Fig. A6, respectively. We observe there is a consistent gap between the values with and without IMMA, which indicates that IMMA prevents the model from learning the personalized content effectively. We also show the results of target and other concepts in Fig. A7, Fig. A8, and Fig. A9. The gap between the two lines shows IMMA immunized the pre-trained model from the target concept while maintaining the ability to be fine-tuned and generate images of other concepts.

A1.3 IMMA for multiple concepts

We conducted experiments of immunization on multiple concepts by running IMMA on each target concept sequentially. As shown in Fig. A10 and Fig. A11, after running IMMA three times on three target concepts: castle, chair and guitar, the immunized model can protect the model from using Textual Inversion to learn any of the concepts. Thus, IMMA has the potential to be extended to multiple-concept scenarios. However, the similarities of other concepts may drop more than the single-concept case, which is worth further research.

A2 Additional qualitative results

A2.1 Comparison with MIST

In Fig. A12, we show additional results of MIST [27]. Textual Inversion using images noised by MIST fails to learn the concept (3^{rd} column). However, after compressing the MISTed images with JPEG, such protection disappears (4^{th} column). Finally, we show that MIST fails to protect personalization items against the adaptation of DreamBooth (5^{th} column). We followed the default parameters of MIST using the strength of the adversarial attack being 16 and the iterations of the attack being 100.

A2.2 IMMA on preventing cloning of face images

We now show that IMMA can effectively restrict the model from duplicating face images of a particular person. We conduct experiments using the datasets from Kumari et al. [24] which contain face images. We use *"a photo of [V] person"* as the prompt. From Fig. A13, we observe that after implementing IMMA on the target person, the model loses its capacity to generate images of that identity using Dreambooth LoRA.

A2.3 IMMA on datasets from Dreambooth [39] and Textual Inversion [12] directly reported in their paper

For the four sets of images shown in Fig. A14, we follow the prompts provided in the corresponding papers. The upper block shows the results of Dreambooth, and the lower block shows the results of Textual Inversion. As we can see, the generation with IMMA successfully prevents the model from generating content of target concepts.

A2.4 Visualization of generation with negative metrics in Tab. 4

In Tab. 4, there are two datasets shown with negative evaluation metric values. To study this, we provide the qualitative results for those corresponding datasets in Fig. A15. In both cases, we observe that DreamBooth *failed to learn the target concept* even without using IMMA.

A3 User Study

The user study is designed to evaluate the generation quality and similarity to the reference images after adaptation with and without IMMA. The question includes both relearning erased styles and personalization adaptation.

Re-learning artistic styles. We evaluate IMMA on preventing style relearning with erased models on eight artistic styles. For each style, the participants are shown four reference images randomly selected from the training images of that artist, *i.e.*, the images generated by SD V1-4 conditioned on *"an artwork of {artist}"*. We provide two images per question for participants to choose from, generation with and without IMMA, respectively. The judgment criteria are image quality (reality) and similarity to reference images. We provide the interface of the user study in Fig. A16.

Personalization adaptation. We also evaluate IMMA on personalization adaptation. For each personal item, the participants are shown four reference images that serve as training images for adaptation. We provide two images per question for participants to choose from, generation with and without IMMA, respectively. The judgment criteria are image quality (reality) and similarity to reference images.

User study for MIST. To evaluate the effect of MIST, we conducted a user study for MIST on personalization adaptation. The setting is identical to that of IMMA except that one of the images to choose from is the generation with MIST instead of IMMA.

A4 Additional Experimental Details

We build our codes on the example code from Diffusers (https://github.com/huggingface/diffusers/tree/main/examples). The pre-trained diffusion

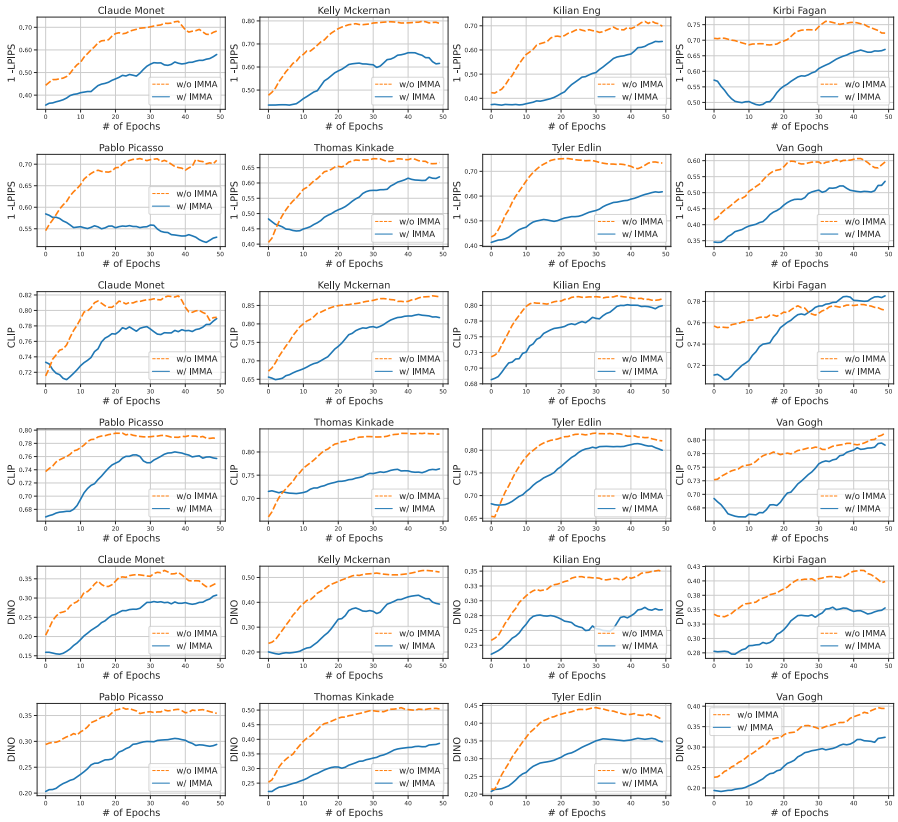

Fig. A2: LPIPS, CLIP, and DINO of LoRA on artistic style erased model Adaptation w/o and w/ IMMA.

model is downloaded from the checkpoint of Stable Diffusion V1-4(https://huggingface.co/CompVis/stable-diffusion-v1-4). Please refer to README.md of our attached code for hyperparameters and experimentation instructions. Note that we use the same set of hyperparameters for each adaptation method across all datasets.

Backbone model for evaluation. We use 'ViT-B/32' for CLIP, 'ViT-S/16' for DINO and AlexNet for LPIPS.

Datasets. We collected our datasets from the following sources: (i) ImgaeNet [9] as in ESD [13] for object relearning. (ii) Eight artistic styles as in ESD [13] for style relearning. (iii) CustomConcept101 [24] for personalization adaptation.

Run time and memory consumption. The running time and memory consumption for IMMA on a specific fine-tuning algorithm $\mathcal{A}$ are comparable with adapting $\mathcal{A}$ on the pre-trained models, *e.g.*, training IMMA against concept relearning with LoRA takes 6 minutes and 15GB GPU memory usage in one Nvidia A30 GPU, where the training step is 1000 with a batch of one.

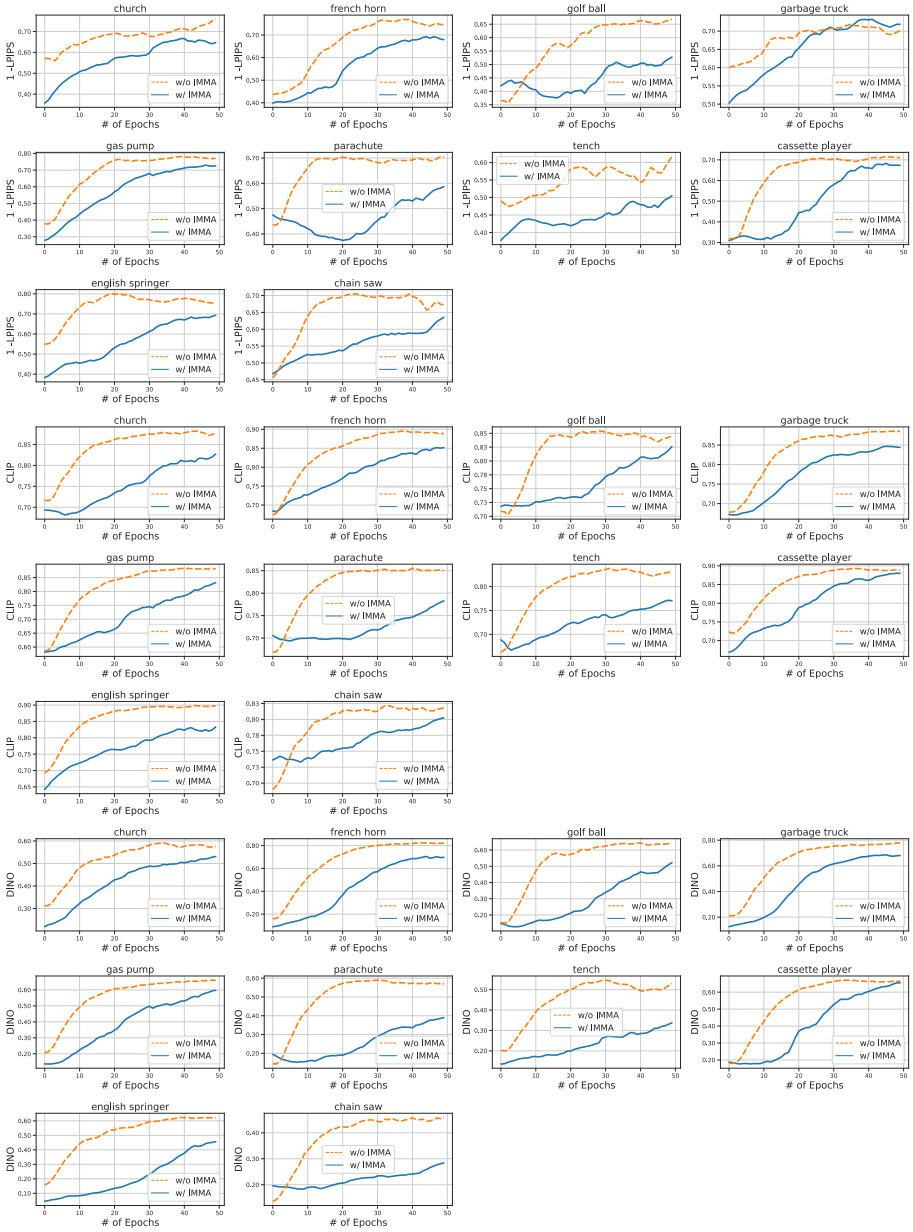

Fig. A3: LPIPS, CLIP, and DINO of LoRA on object erased model Adaptation w/o and w/ IMMA. Our method can prevent models from generating images with target concepts and good quality, as indicated by the high LPIPS, and low CLIP scores.

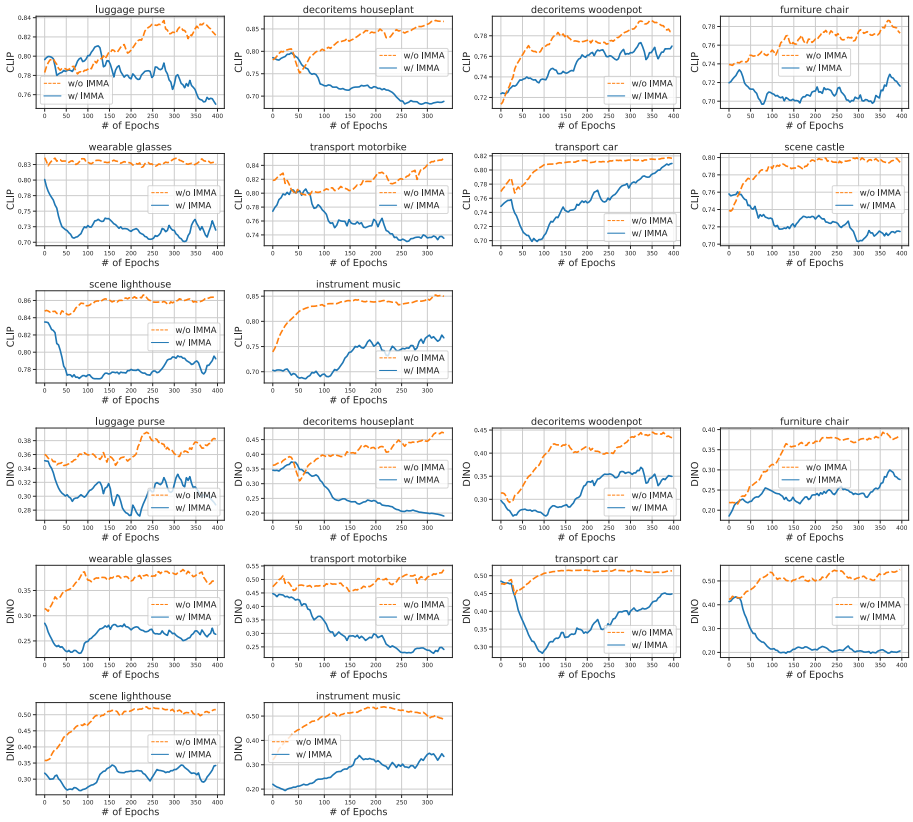

Fig. A4: CLIP and DINO versus reference images after Textual Inversion Adaptation w/o and w/ IMMA.

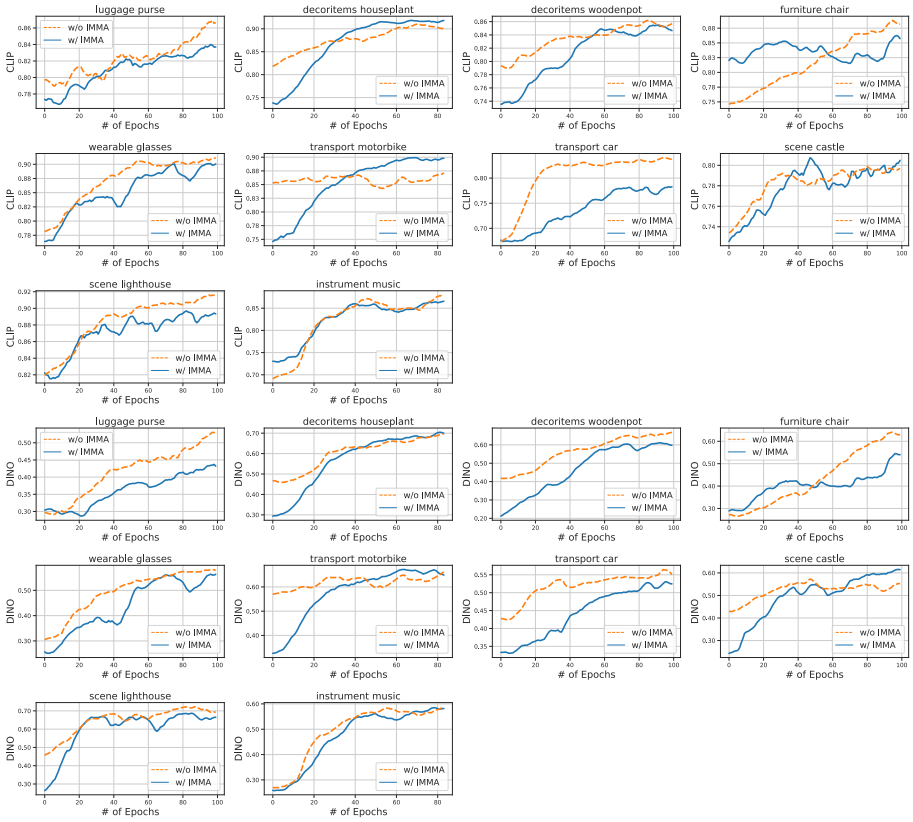

Fig. A5: CLIP and DINO versus reference images after Dreambooth Adaptation w/o and w/ IMMA.

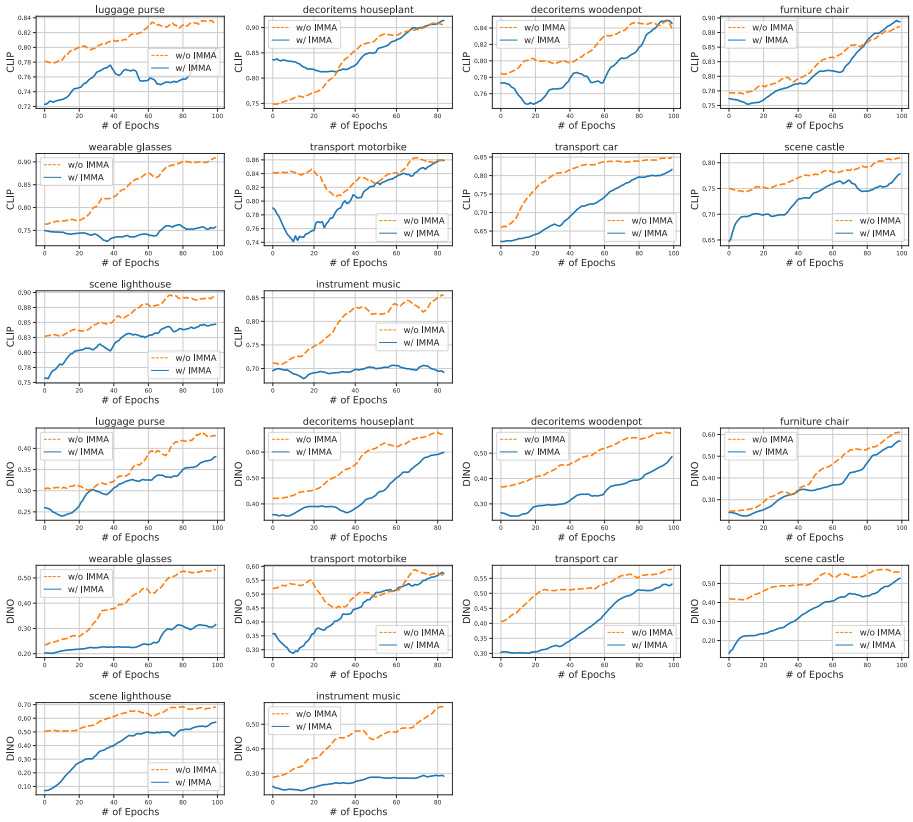

Fig. A6: CLIP and DINO versus reference images after Dreambooth LoRA Adaptation w/o and w/ IMMA.

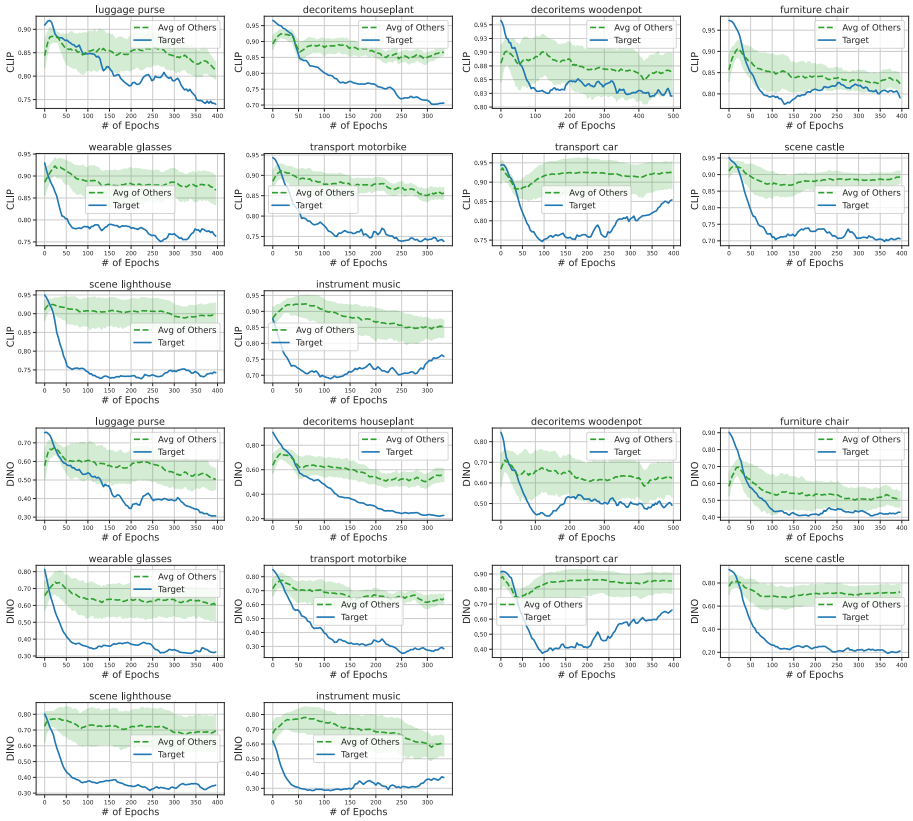

Fig. A7: CLIP and DINO of Textual Inversion Adaptation w/o and w/ IMMA.

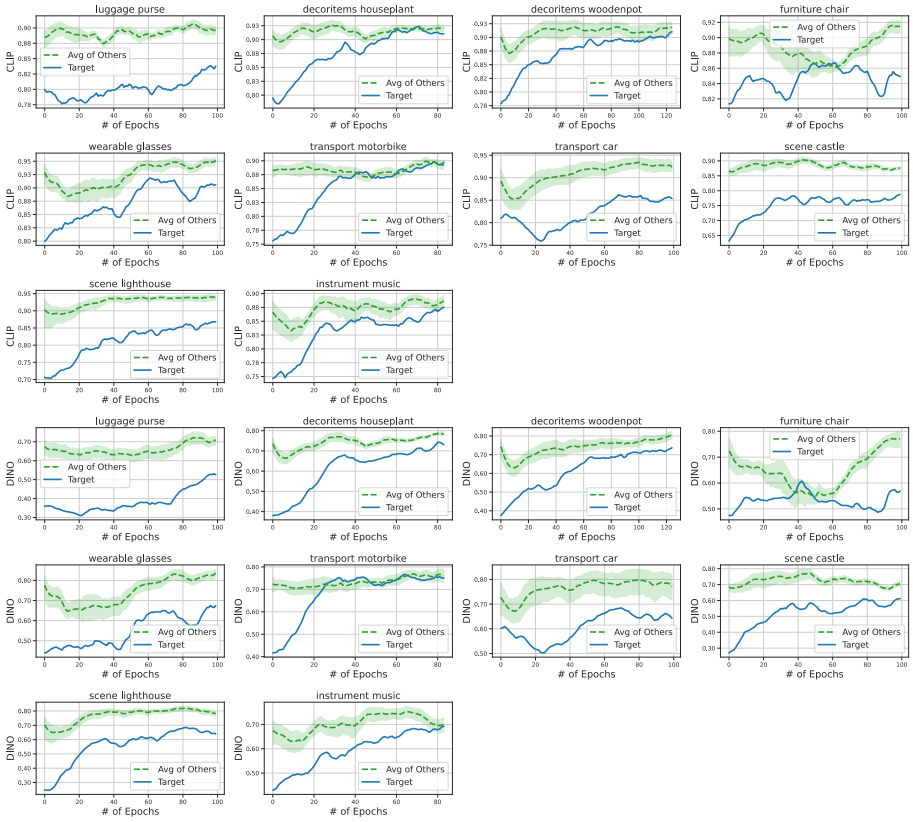

Fig. A8: CLIP and DINO of Dreambooth Adaptation w/o and w/ IMMA.

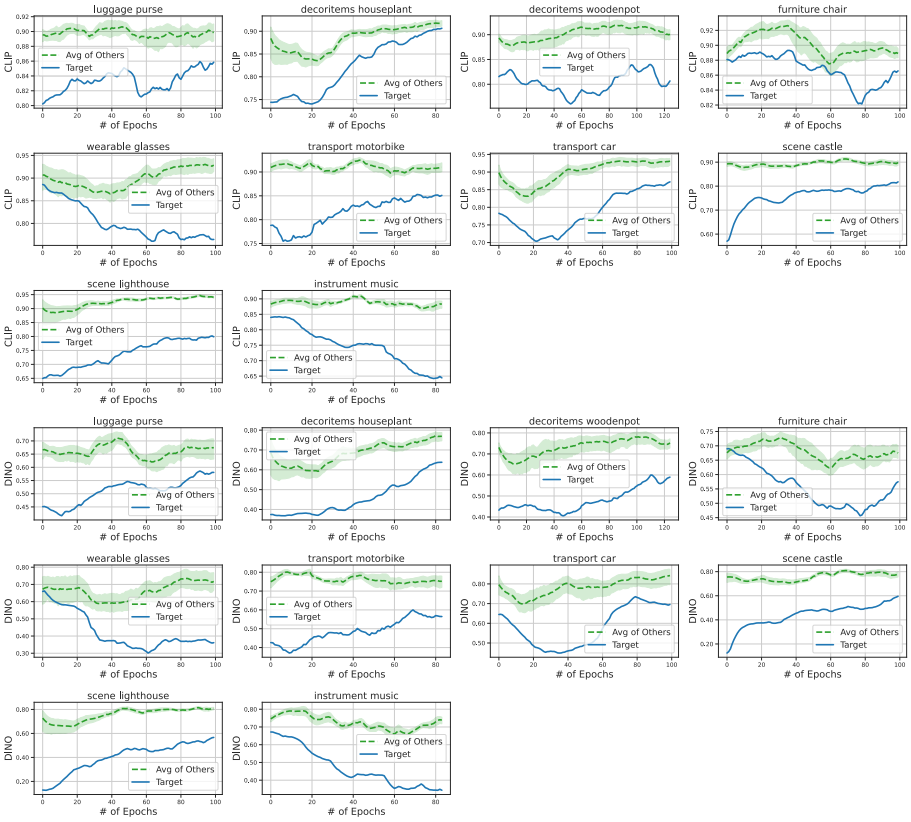

Fig. A9: CLIP and DINO of Dreambooth LoRA Adaptation w/o and w/ IMMA.

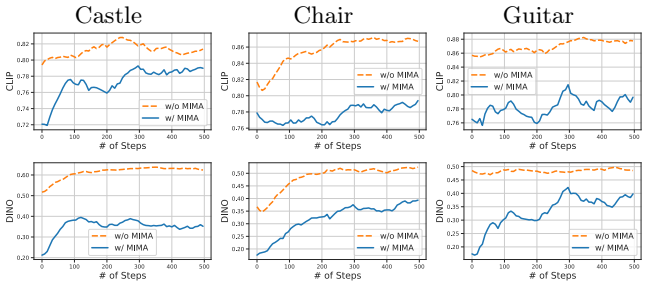

Fig. A10: CLIP and DINO of Textual Inversion w/o and w/ IMMA on multiple concepts.

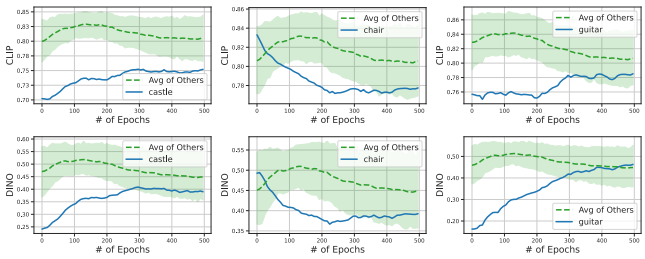

Fig. A11: CLIP and DINO similarity on other concept *vs*. target concept. The adaptation method is Textual Inversion. The gap between the two lines shows `RSGR`.

Fig. A12: Additional results on MIST. We observe that MIST successfully prevented personalization against Textual Inversion. However, MIST is unsuccessful when JPEG is applied to the image, as reported in their paper, or when DreamBooth is used for adaptation.

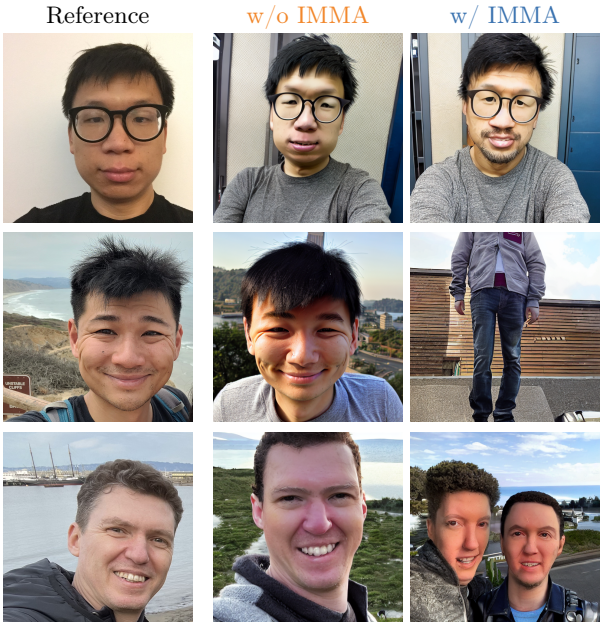

Fig. A13: IMMA on celebrities with adaptation of DreamBooth LoRA.

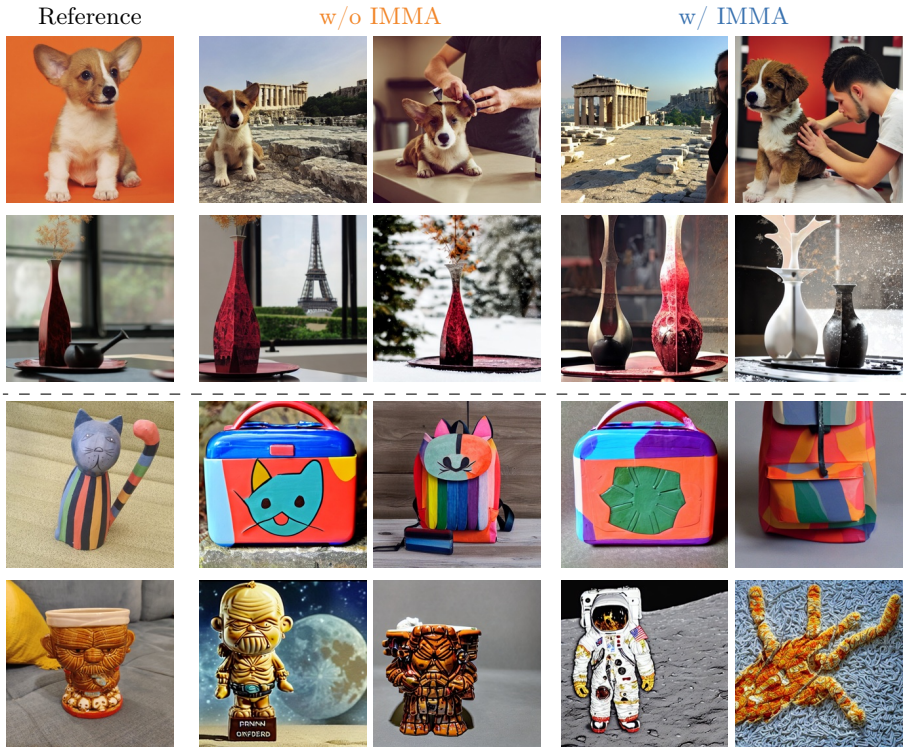

Fig. A14: Results of IMMA on Dreambooth (upper block) and Textual Inversion (lower block) datasets.

Fig. A15: Generation of datasets with negative metric values in Tab. 4. We observe that the base adaptation of DreamBooth's personalization adaptation failed even without IMMA.

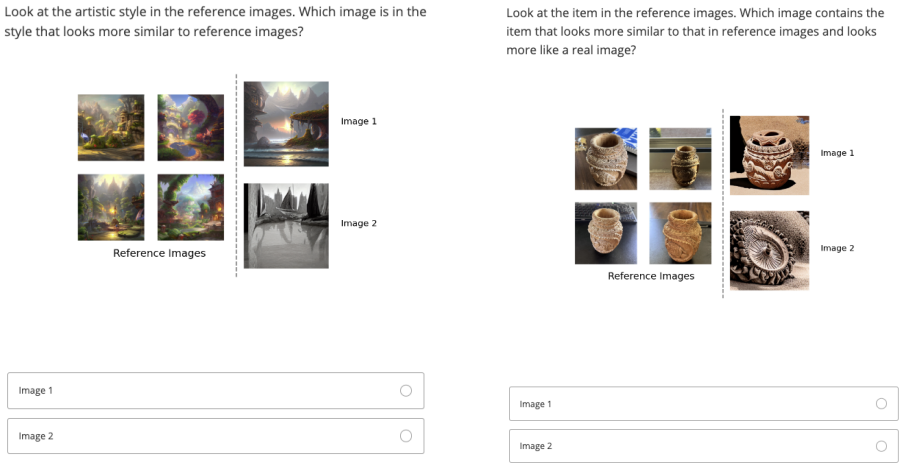

Fig. A16: Illustration of our user study survey. We show four reference images to the user and ask them to select images generated by different methods for comparison.

References

1. Bedapudi, P.: NudeNet: Neural nets for nudity classification, detection and selective censoring. https://github.com/platelminto/NudeNetClassifier (2019)
2. Bengio, Y.: Gradient-based optimization of hyperparameters. Neural Computation (2000)
3. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. In: Proc. ACML (2011)
4. Bird, C., Ungless, E., Kasirzadeh, A.: Typology of risks of generative text-to-image models. In: Proc. AIES (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. CVPR (2021)
6. CreativeML Open RAIL-M: Stable Diffusion LICENSE File (2023), URL https://github.com/CompVis/stable-diffusion/blob/main/LICENSE
7. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D., Li, K., Zhao, Y., Petrovic, V., Singh, M.K., Motwani, S., Wen, Y., Song, Y., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P., Parikh, D.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
8. DeepFloyd Lab at StabilityAI: DeepFloyd IF. https://github.com/deep-floyd/IF (2023)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR (2009)
10. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Proc. NeurIPS (2021)
11. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proc. ICML (2017)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Proc. ICLR (2023)
13. Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proc. ICCV (2023)
14. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proc. WACV (2024)
15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proc. ICLR (2015)
16. Harwell, D.: AI-generated child sex images spawn new nightmare for the web. The Washington Post (2023), URL https://www.washingtonpost.com/technology/2023/06/19/artificial-intelligence-child-sex-abuse-images/
17. Heikkilä, M.: This artist is dominating ai-generated art. and he's not happy about it. MIT Technology Review. Retrieved March **16**, 2023 (2022)

18. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. arXiv preprint arXiv:2305.10120 (2023)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proc. NeurIPS (2020)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proc. ICLR (2022)
21. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: Proc. NeurIPS (2021)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. ICLR (2015)
23. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proc. ICCV (2023)
24. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proc. CVPR (2023)
25. Larsen, J., Hansen, L.K., Svarer, C., Ohlsson, M.: Design and regularization of neural networks: the optimal use of a validation set. In: IEEE Signal Processing Society Workshop (1996)
26. Liang, C., Wu, X.: Mist: Towards improved adversarial examples for diffusion models (2023)
27. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In: Proc. ICML (2023)
28. Lorraine, J., Vicol, P., Duvenaud, D.: Optimizing millions of hyperparameters by implicit differentiation. In: Proc. AISTATS (2020)
29. Mei, S., Zhu, X.: Using machine teaching to identify optimal training-set attacks on machine learners. In: Proc. AAAI (2015)
30. Moore, S.: Can the law prevent AI from duplicating actors? It's complicated. Forbes (Jul 2023), URL https://www.forbes.com/sites/schuylermoore/2023/07/13/protecting-celebrities-including-all-actors-from-ai-with-the-right-of-publicity/?sh=5c56ba4159ec
31. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proc. ICML (2022)
32. Noveck, J., O'Brien, M.: Visual artists sue ai companies in sf federal court for repurposing their work. Associated Press (2023)
33. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
34. Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., Zhang, Y.: Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. arXiv preprint arXiv:2305.13873 (2023)
35. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)

36. Rando, J., Paleka, D., Lindner, D., Heim, L., Tramèr, F.: Red-teaming the stable diffusion safety filter. In: Proc. NeurIPS ML Safety Workshop (2022)
37. Ren, Z., Yeh, R., Schwing, A.: Not all unlabeled data are equal: Learning to weight data in semi-supervised learning (2020)
38. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. CVPR (2022)
39. 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: Proc. CVPR (2023)
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Proc. NeurIPS (2022)
41. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious AI-powered image editing. In: Proc. ICML (2023)
42. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proc. CVPR (2023)
43. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models. In: Proc. NeurIPS (2022)
44. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models. In: USENIX Security Symposium (2023)
45. SmithMano: Tutorial: How to remove the safety filter in 5 seconds. Reddit (2022)
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proc. ICML (2015)
47. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Proc. NeurIPS (2019)
48. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation (2011)
49. Wang, Z., Chen, C., Liu, Y., Lyu, L., Metaxas, D., Ma, S.: How to detect unauthorized data usages in text-to-image diffusion models. arXiv preprint arXiv:2307.03108 (2023)
50. Yeh, R.A., Hu, Y.T., Hasegawa-Johnson, M., Schwing, A.: Equivariance discovery by learned parameter-sharing. In: Proc. AISTATS (2022)
51. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2211.08332 (2023)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. CVPR (2018)

53. Zhao, Z., Duan, J., Hu, X., Xu, K., Wang, C., Zhang, R., Du, Z., Guo, Q., Chen, Y.: Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. arXiv preprint arXiv:2306.01902 (2023)
54. Zheng, A.Y., He, T., Qiu, Y., Wang, M., Wipf, D.: Graph machine learning through the lens of bilevel optimization. In: Proc. AISTATS (2024)
55. Zheng, A.Y., Yang, C.A., Yeh, R.A.: Learning to obstruct few-shot image classification over restricted classes. In: Proc. ECCV (2024)