©PLUG-IN AUTHORIZATION FOR HUMAN CONTENT COPYRIGHT PROTECTION IN TEXT-TO-IMAGE MODEL

Chao Zhou [1], Huishuai Zhang [2], Jiang Bian [3], Weiming Zhang [1], Nenghai Yu [1]
[1]University of Science and Technology of China, [2]Peking University, [3]Microsoft Research
zc1696190340@mail.ustc.edu.cn, zhanghuishuai@pku.edu.cn
jiabia@microsoft.com, {zhangwm,ynh}@ustc.edu.cn

ABSTRACT

This paper addresses the contentious issue of copyright infringement in images generated by text-to-image models, sparking debates among AI developers, content creators, and legal entities. State-of-the-art models create high-quality content without crediting original creators, causing concern in the artistic community. To mitigate this, we propose the ©Plug-in Authorization framework, introducing three operations: addition, extraction, and combination. Addition involves training a ©plug-in for specific copyright, facilitating proper credit attribution. Extraction allows creators to reclaim copyright from infringing models, and the combination enables users to merge different ©plug-ins. These operations act as permits, incentivizing fair use and providing flexibility in authorization. We present innovative approaches, "Reverse LoRA" for extraction and "EasyMerge" for seamless combination. Experiments in artist-style replication and cartoon IP recreation demonstrate ©plug-ins' effectiveness, offering a valuable solution for human copyright protection in the age of generative AIs.

1 INTRODUCTION

Large foundation models (Brown et al.; Touvron et al., 2023) are trained with extensive, high-quality datasets like The Pile (Gao et al., 2020), C4 (Raffel et al., 2020), LAION (Schuhmann et al., 2022) and other enormous undisclosed data sources, which contain copyrighted human contents. At the same time, these models not only excel at generating content based on user prompts (cha; OpenAI, 2023; Rombach et al., 2022b; Ramesh et al., 2021; 2022), but also have the potential of memorizing the exact training data thanks to the huge capacity in their gigantic numbers of parameters (Carlini et al., 2021; 2023a).

Such training of AI models has sparked copyright infringement concerns among content providers, artists, and users. A notable instance is a lawsuit filed by The New York Times against OpenAI and Microsoft (NYT), alleging the unauthorized use of a vast number of articles to train automated chatbots. The lawsuit seeks the destruction of the allegedly infringing chatbots and their associated training data. Similar concerns and legal actions are also emerging in the field of text-to-image generation (law).

In this paper, we reassess the justification for copyright laws, stressing their fundamental role in safeguarding creative expression across diverse mediums. Copyright endeavors to spur scientific and artistic advancements by providing authors with exclusive rights to their creations for a specified duration, fostering both innovation and dissemination. However, modern generative AI models present challenges in fairly compensating copyright holders, affecting artists and contributors on platforms like StackOverflow. This quandary could impede the accessibility of new data for machine learning, underscoring the broader societal ramifications of addressing copyright concerns amid the rise of advanced generative models.

To address the challenges of copyright in generative models, we present the "©Plug-in Authorization" framework (see Figure 1), aligning with existing Intellectual Property (IP) management practices. Base model providers, like Stability AI, serve as repositories for copyright plug-ins, where artists can register their works and receive rewards for usage.

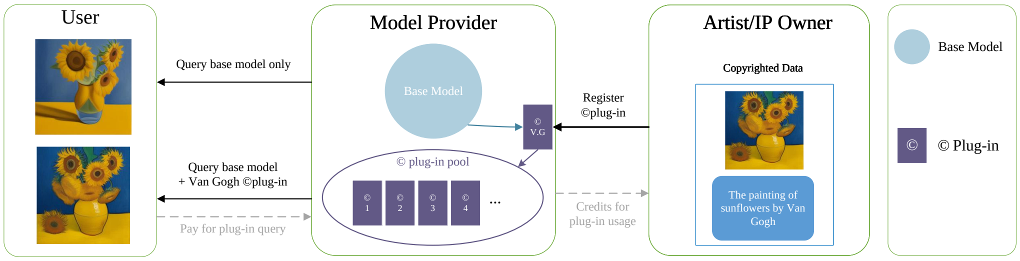


Figure 1: **©Plug-in Authorization Framework.** This process involves three entities: model providers, IP owners (artists), and users. Model providers offer services, track plug-in usage, and distribute rewards to IP owners. IP owners register their plug-ins via *addition* or *extraction*. Users obtain authorization through relevant plug-ins to generate copyrighted images. These plug-ins constitute a pool accessible to users for generating authorized content.

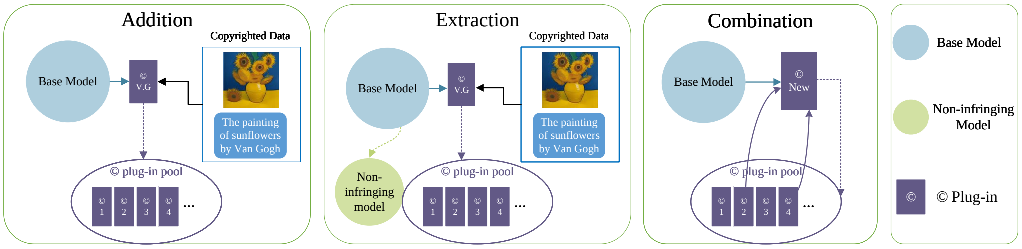


Figure 2: **Three foundational operations achieving ©Plug-in Authorization.** *Addition* creates a plug-in when the copyrighted work is new to the base model. Conversely, *extraction* generates a plug-in from the base model if the work is already present. Subsequently, the *combination* operation merges multiple plug-ins, allowing the generation of diverse concepts while maintaining a non-infringing model.

Technically to enable an effective and efficient copyright authorization, the plug-ins, as permits, should be easily created by *addition* if copyrighted works are new to the base models, or by *extraction* if the copyrighted works are already infringed by the base model. Moreover, the plug-ins should be easily *combined*, which allows copyright holders to merge multiple plug-ins into a new one or enables end users to generate images with multiple copyrighted works. Meanwhile, for efficient execution, these operations should be implemented as light adaptations to the base model, e.g., parameter-efficient tuning methods or prompt designs.

In this paper, we introduce three foundational operations - *addition, extraction*, and *combination* - implemented using the Low-Rank Adaptor (LoRA) method (Hu et al., 2022). These operations are essential for realizing the ©Plug-in Authorization (see Figure 2 for an overview).

Civitai (civ) commendably implements the *addition* operation, allowing users to train and share LoRA components for generating specific figures. However, the *extraction* and *combination* operations pose significant challenges and are not publicly available in Civitai.

The *extraction* operation aims to separate the generative model into a non-infringing base model and copyrighted plug-ins. Traditional methods often involve costly retraining using only non-infringing data and applying LoRA with copyrighted data, which is impractical. This paper proposes a "Reverse LoRA" approach, initially tuning the model on the target concept and then negating the weights for concept destruction. Subsequent fine-tuning on the surrounding context restores the non-infringing model's contextual generation ability, and the process concludes by reversing the LoRA to obtain the copyright plug-in.

The *combination* operation merges multiple copyrighted plug-ins, which, when added together directly, may produce unpredictable outcomes due to inter-correlations. This paper introduces

"EasyMerge", a method using "data-free layer-wise distillation" for combination. Inspired by conditional generation in generative models, we employ a LoRA component to learn layer-wise outputs of copyright plug-ins under corresponding conditions. This enables the LoRA component to mimic their behavior, facilitating effective combination.

2 ©Plug-in Authorization with Addition, Extraction and Combination

To instantiate the "©Plug-in Authorization" as detailed in the Introduction, we integrate three foundational operations into SDM (Stable Diffusion Model (Rombach et al., 2022a)): *addition*, allowing copyright owners to add a plug-in for their works; *extraction*, enabling owners to extract a plug-in from an infringing base model, and *combination*, allowing users to merge plug-ins for multiple copyrighted concepts. The addition operation utilizes LoRA components added to SDM's attention matrices, learning them with copyrighted data. While the addition is available in existing model-sharing platforms, we delve into the extraction and combination operations in the following sections.

2.1 Extraction: Reverse LoRA

In the pursuit of achieving *extraction*, we introduce a method termed "Reverse LoRA". This approach comprises two key steps to effectively capture the target copyrighted concept without losing the capability of contextual generation. The first step involves the de-concept of the target concept, followed by the subsequent re-context of the surrounding semantics context. To elucidate the process of extracting the target concept "Picasso", Figure 3 provides a visual representation of these two steps. Due to the page limitation, more details can refer to Appendix B.

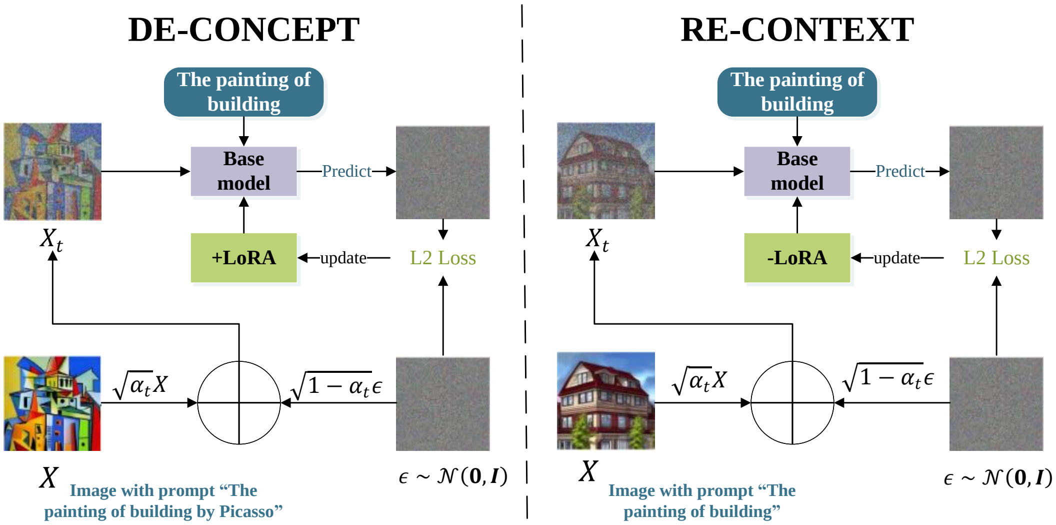

DE-CONCEPT
The painting of building
Base model
Predict
X_t
+LoRA
update
L2 Loss
$\sqrt{\alpha_t}X$
$\sqrt{1-\alpha_t}\epsilon$
X
Image with prompt "The painting of building by Picasso"
$\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
RE-CONTEXT
The painting of building
Base model
Predict
X_t
-LoRA
update
L2 Loss
$\sqrt{\alpha_t}X$
$\sqrt{1-\alpha_t}\epsilon$
X
Image with prompt "The painting of building"
$\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$

Figure 3: **The method of *extraction*** consists of two steps: De-concept and Re-context. The de-concept phase tries to capture the target concept "Picasso" by tuning the LoRA component to match copyrighted images with the contextual prompt "The painting of building". In the re-context stage, we reverse the LoRA (so that successfully forget "Picasso") and then further tune the LoRA with surrounding contextual prompt and image pairs, to ensure the capabilities of generating surrounding context.

2.2 Combination: EasyMerge

When seeking to generate an image featuring both "Snoopy" and "Mickey" concepts, combining existing plug-ins is crucial. However, directly adding these plug-ins may lead to unpredictable outcomes due to inherent correlations. To address this, we propose "EasyMerge", a data-free layer-wise distillation method. This approach involves associating prompts containing distinct copyright information with their respective copyright components during the inference process like the following objective:

$$\arg\min_{w_L} \sum_{k\in S, j\in S_L} \mathbb{E}_{\epsilon,t}\|\phi^j_{w-w_L}(\epsilon, c_k, t) - \phi^j_{w-w_{L_k}}(\epsilon, c_k, t)\|^2, \tag{1}$$

where S is the set of text prompts, S_L is the set of layers equipped with LoRA components, and ϕ^j is the output of layer j's LoRA component. w denotes the base model parameter, w_L denotes the combined plug-in, c_k denotes the prompt k, w_{L_k} is the plug-in of context c_k, ϵ is initial noise and t is the sampling timestep of the diffusion process. Algorithm 1 depicts the concrete steps of optimizing the objective as Equation 1.

Algorithm 1 Combination: EasyMerge method

```
Input: A set $S$ of indices of plug-ins to be combined, base model $w$, diffusion step $T$
Output: Combined LoRA $w_L$
repeat
    for $w_{L_i}, c_i \in S$ do
        $t \sim$ Uniform$([1...T])$; $\epsilon \sim N(0,1)$
        AddHook($w_{L_i}$)    ▷ Capture input $I^j_{w_{L_i}}$ and output $O^j_{w_{L_i}}$ for each layer $j$
        $I^j_{w_{L_i}}, O^j_{w_{L_i}} \leftarrow \Phi_{w+w_{L_i}}(\epsilon, c_i, t)$    ▷ Denoise to obtain $I^j_{w_{L_i}}$ and $O^j_{w_{L_i}}$
        $O^j_{w_L} \leftarrow \phi^j_{w_L}(I^j_{w_{L_i}})$    ▷ Get layer-output $O^j_{w_L}$
        $\mathcal{L} \leftarrow \sum_{j\in S_L} \|O^j_{w_L} - O^j_{w_{L_i}}\|$
        $w_L \leftarrow w_L - \nabla_{w_L}\mathcal{L}$
    end for
until convergence
```

3 Experiments to Verify Efficacy of Operations

Although the main contribution is the copyright authorization framework, we still would like to verify the efficacy of the operations in practice. As the *addition* operation has been well demonstrated by the public, we focus on evaluating *extraction* and *combination* operations. We choose two typical scenarios of copyright infringement: artist-style replication and cartoon IP recreation. We compare our method with the concept ablation approach (Kumari et al., 2023) and Erased Stable Diffusion (ESD) (Gandikota et al., 2023).

Due to the page limitation, we just show the results on cartoon IP recreation. The experiment setup and metric are put in Appendix D.1. The results on artist-style replication are put in Appendix D.2.

3.1 Extraction and Combination of Cartoon IPs

For IP recreation, we showcase results of both *extraction* and *combination*. Figure 4 illustrates the outcomes after extracting three IP characters: Mickey, R2D2, and Snoopy. Images within the red boxes are generated using prompts representing the target IP by the non-infringing model post-extraction. These images notably differ from those produced by the base model, making it challenging to discern the character IP. Conversely, images outside the red boxes representing surrounding IPs maintain similarity to base model outputs.

Our extraction method excels, isolating designated IPs without disrupting the generation of others. Table 1 quantifies its effectiveness, showing approximately a 2.6 times increase in the KID metric for the target IP while keeping surrounding IPs relatively unchanged. Extensive large-scale experiments further confirm that extraction has no impact on generating ordinary items, nor does the extracted model adversely affect routine use (see Appendix D.3).

Additionally, we demonstrate the application of combined copyright plug-ins in single images (Figure 5). The first image, post-extraction of Mickey Mouse and Darth Vader, obscures the IPs. Subsequent images, post-integrating Mickey and Vader plug-ins individually, exclusively feature their corresponding IPs. Finally, adding the combined plug-in restores the model's ability to generate both

IPs in one image. This indicates that after the *combination* step, the non-infringing model's capability to generate either Mickey Mouse or Darth Vader-themed images is deactivated. Upon adding the corresponding ©Plug-in, the model regains this ability, exclusively within each IP's domain.

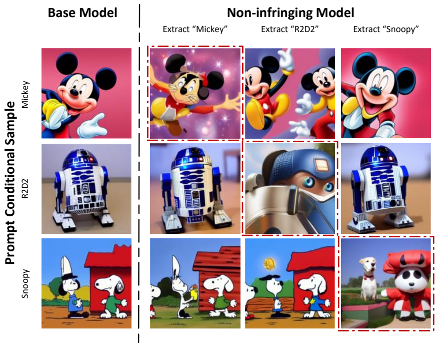


Figure 4: **Results of *extraction* in IP recreation.** Each column of images is generated by a distinct non-infringing model. We can extract any individual IP of Mickey, R2D2, and Vader without affecting the generation of other IPs.

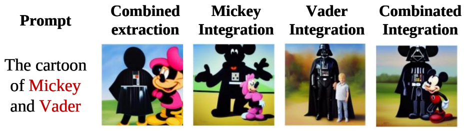


Figure 5: **IP Recreation in a single image.** We can integrate ©plug-in into the non-infringing model to generate either Mickey or Vader in a single image or integrate the combined ©plug-in to generate both of them.

Table 1: **Quantitative comparison in IP recreation.** We increase the KID of the target IP about 2.6 times compared with Concepts-Ablation, while keeping the KID of the surrounding IP on par.

Metrics	Methods	Target IP ↑	Surrounding IP ↓
KID×10^3	Extraction (Ours)	**131**	17
	Concepts-Ablation	50	15

4 Discussion, Open Questions and Limitations

The rapid advancement of generative AI models has sparked significant concern over their potential to produce copyright-infringing content. As these models elevate the quality of their outputs, the lack of proper attribution to human creators becomes increasingly problematic. To mitigate these societal concerns, we introduce the "©Plug-in Authorization" framework, a system designed to align with the principles of copyright law and ensure that the use of copyrighted material in AI-generated content is appropriately managed and rewarded.

Our framework leverages the Low-Rank Adaptor (LoRA) method to integrate copyrighted data into plug-ins, facilitating the tracking of content usage and the fair distribution of rewards to original creators. This approach not only respects the rights of content creators but also provides a clear and accessible mechanism for users to engage with copyrighted material in a legally compliant manner.

However, the framework faces challenges in the efficient management of a vast array of plug-ins. As the number of plug-ins grows, maintaining a user-friendly interface and ensuring seamless access to specific generations becomes critical. This requires innovative solutions to organize and retrieve plug-ins effectively, without compromising the user experience.

Another challenge is the updating of the base model. Retraining the entire collection of plug-ins following a model update can be prohibitively expensive and time-consuming. Ensuring backward compatibility, so that previously created plug-ins continue to function with new model iterations, is essential to avoid obsolescence and maintain user trust.

In conclusion, our ©Plug-in Authorization framework represents a significant step towards balancing the innovative potential of generative AI with the imperative of copyright protection. By addressing the challenges of plug-in management, model updating, and performance maintenance, we aim to foster an ecosystem where creativity and legal compliance can coexist and thrive.

Acknowledgement

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62372423.

References

NYT Complaint Dec2023. https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

ChatGPT. https://chat.openai.com/.

Civitai. https://civitai.com/.

AI Art Generators Spark Multiple Copyright Lawsuits. https://www.hollywoodreporter.com.

Ryan Abbott and Elizabeth Rothman. Disrupting creativity: Copyright law in the age of generative artificial intelligence. *Florida Law Review, Forthcoming*, 2022.

Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition*, pp. 18511–18521, 2022.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Olivier Bousquet, Roi Livni, and Shay Moran. Synthetic data generators–sequential and private. *Advances in Neural Information Processing Systems*, 33:7114–7124, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pp. 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *International Conference on Learning Representations*, 2023a.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023b.

Niva Elkin-Koren, Uri Hacohen, Roi Livni, and Shay Moran. Can copyright be reduced to privacy? *arXiv preprint arXiv:2305.14822*, 2023.

Gongfan Fang, Jie Song, Chengchao Shen, Xinchao Wang, Da Chen, and Mingli Song. Data-free adversarial distillation. *arXiv preprint arXiv:1912.11006*, 2019.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Annual ACM SIGACT Symposium on Theory of Computing*, 2020.

Giorgio Franceschelli and Mirco Musolesi. Copyright in generative deep learning. *Data & Policy*, 4:e17, 2022.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

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

Hanlin Li, Brent Hecht, and Stevie Chancellor. All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pp. 584–595, 2022a.

Hanlin Li, Brent Hecht, and Stevie Chancellor. Measuring the monetary value of online volunteer work. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pp. 596–606, 2022b.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pp. 20852–20867. PMLR, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717, 2022.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

OpenAI. Dall·E 3. https://openai.com/dall-e-3, 2023.

Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramer. Data poisoning won't save you from facial recognition. In *International Conference on Learning Representations*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. Red-teaming the stable diffusion safety filter. In *NeurIPS ML Safety Workshop*, 2022.

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

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*, pp. 10684–10695, 2022b.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4323–4332, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Anton Troynikov. Stable Attribution. https://www.stableattribution.com, 2023.

Nicholas Vincent and Brent Hecht. A deeper investigation of the importance of wikipedia links to search engine results. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–15, 2021.

Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 215–227, 2021.

Nikhil Vyas, Sham Kakade, and Boaz Barak. Provable copyright protection for generative models. *arXiv preprint arXiv:2302.10870*, 2023.

Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Haonan Zhong, Jiamin Chang, Ziyue Yang, Tingmin Wu, Pathum Chamikara Mahawaga Arachchige, Chehara Pathmabandu, and Minhui Xue. Copyright protection and accountability of generative ai: Attack, watermarking and attribution. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 94–98, 2023.

A Related Work

In order to position our work in the vast literature, we review related work through two perspectives: scope and technique. It is worthy to note that some of the literature touch both sides and we organize them in a way most related to ours.

A.1 Scope Related: Copyright, Data Contribution and Credit Attribution

Recent text-to-image generative models are trained with large scale datasets (Schuhmann et al., 2022; Liu et al., 2022), which cannot be guaranteed free of copyrighted data. At the same time, the state-of-the-art models are capable of generating high-quality and valuable creative images comparable to human creators or even memorizing the data points in the training set (Carlini et al., 2023b), which arouses copyright concerns about the training data and brings anxiety to the artist community.

Numerous efforts have been made for copyright protection of training data (Zhong et al., 2023). A direct approach is removing the copyrighted images from the training set, which may involve cumbersome cost due to the size of the training sets and may significantly degrade the model performance (Feldman, 2020). Another direct approach is post filtering, refusing to generate images with copyrighted concepts, e.g., Schramowski et al. (2023) proposes *Safe Latent Diffusion* to guide latent representation away from target concepts in the inference process, which nonetheless can be bypassed by a user with access to the model (Rando et al., 2022). As an example, OpenAI Dall·E3 (OpenAI, 2023) declines requests for generating an image in the style of a living artist and promises that creators can also opt their images out from training of future image generation models. Many papers discuss the idea of concept removal, which will be reviewed in later section.

Shan et al. (2023) propose *Image Cloaking* that suggests adding adversarial perturbations before posting artistic works on the internet so as to make them unlearnable for machine learning model, which has been pointed out to be hard to defend against future learning algorithms (Radiya-Dixit et al., 2021).

Theoretically, Bousquet et al. (2020); Elkin-Koren et al. (2023) connect the copyright protection of training data with the concept of differential privacy and discuss their subtle differences. Vyas et al. (2023) further formulate the copyright problem with a *near free access* (NAF) notion to bound the distance of the generative distributions of the models trained with and without the copyrighted data.

Our paper distinguishes largely from all previous works as we do not try to prohibit generating copyrighted concepts but instead we introduce a copyright authorization for the generative model to reward the copyright owners with fairness and transparency. From this aspect, our paper is also related with literature of monetizing the training data (Vincent & Hecht, 2021; Vincent et al., 2021; Li et al., 2022b;a) or attributing credits for the generative contents (Troynikov, 2023), but we establish a very distinct way to reward the authorship.

Heated discussion is also around the copyright for AI generated art work Franceschelli & Musolesi (2022); Abbott & Rothman (2022). The Review Board of the United States Copyright Office has recently refused the copyright registration of a two-dimensional AI generated artwork entitled "A Recent Entrance to Paradise". However, Abbott & Rothman (2022) argues for giving the copyright to AI generated works, which will encourage people to develop and use creative AI, promote transparency and eventually benefit the public interest.

A.2 Technique Related: Concept Removal, and Negative Sampling

Our *extraction* operation is closely related with the *concept removal* for generative models. Gandikota et al. (2023); Kumari et al. (2023) remove target concepts by matching the generation distribution of contexts with target concepts and that of contexts without target concepts. Zhang et al. (2023) forget target concepts by minimizing the cross attention of target concepts with that of target images. Heng & Soh (2023) leverage the reverse process of continual learning to promote the controllable forgetting of target contents in deep generative models.

We note that negative sampling (Ho & Salimans, 2022) can also prevent generating certain concepts. Specifically, end users can set conditional context and negative context to guide the diffusion process to generate images conforming the conditional context while being far away from the negative context. Only negative sampling cannot stop copyright infringing generation because the contexts are set freely and adversarially by end users.

In contrast, for a specific copyrighted concept, our *extraction* operation takes an "inverse LoRA" approach to disentangle the base model into two part: non-infringing base model and the plug-in LoRA component for copyrighted concept. Specifically, we use negative sampling to generate non-infringing images, which serves as training data for copyright plug-in. From the aspect of parameter efficient fine-tuning, our paper is related with literature (Alaluf et al., 2022; Ruiz et al., 2023; Gal et al., 2022; Hu et al., 2022; Huang et al., 2023).

Our *combination* operation is related with the widely studied "knowledge distillation" (Liang et al., 2023; Lopes et al., 2017; Sun et al., 2019; Hinton et al., 2015; Fang et al., 2019), but entails large difference from previous work. We combine multiple copyright plug-ins that are LoRA components for different targets, and we take data free approach due to practical constraint.

B More Details in Extraction

To elucidate the process of extracting the target concept "Picasso", Figure 3 provides a visual representation of these two steps,*De-concept* step and *Re-context* step.

B.1 Step1: De-concept

Our goal is to extract Picasso-related information from the base model and incorporate it into a copyright plug-in using LoRA (w_L). This involves finding the information representing "Picasso" and capturing model parameter changes during alignment between image generative models familiar and unfamiliar with Picasso.

In practice, we align copyrighted image generation (e.g., images of "the painting of building by Picasso") with the text prompt "the painting of building" on the base infringing model. This alignment is expressed as:

$$\mathbb{E}[\Phi_{(w)}(\epsilon, c^*, t)] = \mathbb{E}[\Phi_{(w+w_L)}(\epsilon, c, t)] \tag{2}$$

where Φ is the denoising function, w denotes the original network parameter, w_L is the LoRA component, c is the prompt "the painting of building",c^* is the prompt "the painting of building by Picasso", ϵ is the initial noise, and t is the sampling timestep.

To achieve Equation 2, we optimize the following objective concerning the LoRA w_L and freezing others,

$$\arg\min_{w_L} \; \mathbb{E}_{\epsilon, X^*, c, t} \|\Phi_{(w+w_L)}(X^*_t, c, t) - \epsilon\|^2 \tag{3}$$

where X^* is the copyrighted image (or generated by the infringing model with the prompt "the painting of building by Picasso"), $X^*_t = \sqrt{\alpha_t} X^* + \sqrt{1-\alpha_t}\epsilon$ is the noisy version of X^*, c is the prompt of "the painting of building", w is the original network and w_L is the LoRA weight.

By adding such a LoRA, the base model can generate Picasso-style images even when the prompts do not contain the word "Picasso". Hence, the LoRA represents the copyrighted Picasso style, and $w - w_L$ would produce a non-infringing model, which can thought to be an analogy of a negative LoRA. However, directly using $w - w_L$ as the non-infringing model diminishes the capability to generate images with surrounding context ("the painting of building"), as shown in Figure 6 in Appendix C. This prompts us to further tune the LoRA with pairs of images and surrounding semantics texts.

B.2 Step2: Re-context

To mitigate the performance degradation of the non-infringing model when generating images with surrounding contextual prompts, we implement a memorization phase post-unlearning. This involves tuning the LoRA component using images and textual caption corresponding to the surrounding prompt "The painting of building". The image dataset is curated by randomly prompting the base model with the contextual prompts "The painting of building", leveraging the negative prompt (Ho & Salimans, 2022) "Picasso" to guide the generation far away from the target concept "Picasso" as much as possible.

In practice, we further optimize w_L with the objective in Equation 4 where X, c represent the constructed (image, prompt) pairs for the recovery of surrounding contextual generation.

$$\arg\min_{w_L} \mathbb{E}_{\epsilon,X,c,t}\|\Phi_{(w-w_L)}(X_t, c, t) - \epsilon\|^2 \tag{4}$$

The model $w - w_L$ cannot generate the images with Picasso style due to the unlearning step but performs well with surrounding prompts owing to the memorization step. Through the *extraction* operation, we derive a non-infringing model $\tilde{w} = w - w_L$ and a ©plug-in w_L. By incorporating the ©plug-in, the model reverts to the base model w, capable of successfully generating the artworks with the "Picasso" style. The intermediate results in *extraction* are visually depicted in Figure 6 in Appendix C, showcasing the successful extraction of the targeted copyright, while maintaining the model's ability to generate images with surrounding context.

C Intermediate results of extraction

We analyze the *extraction* process (see Figure 6), revealing distinct phases in the evolution of the non-infringing model's performance. After de-concept, the model undergoes a transient phase with a temporary impairment in generating semantically meaningful images. Subsequent re-context restores its proficiency in generating semantically rich images. However, the model remains unable to produce artwork in Picasso's distinctive style, indicating the success of the de-concept step. This step effectively removes the target concept, causing temporary impairment. The re-context step resolves this without reintroducing the target concept. In essence, the extraction process successfully achieves targeted concept extraction.

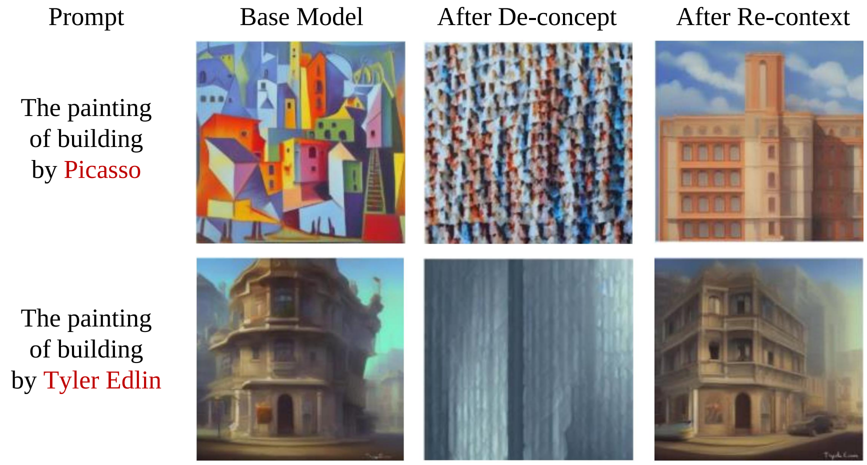

Prompt
Base Model
After De-concept
After Re-context
The painting of building by Picasso
The painting of building by Tyler Edlin

Figure 6: **Intermediate results of *extraction*.** After the de-concept step, the non-infringing model's generative abilities become significantly limited, predominantly manifesting as the production of noise. After the re-context step, the generation's prowess is rejuvenated, but due to the absence of learning Picasso-style images, the model remains unable to generate artwork in the style of Picasso.

D More Experiment

D.1 Experiment Setup, Metrics and Baselines

Experiment Setup For all experiments, we tune the attention part in U-Net construction of Stable Diffusion Model v1.5 (Rombach et al., 2022a), which is observed to generate better images with celebrities or artistic styles than the series of Stable Diffusion Model v2.

Here is how we generate data from the pre-trained model for extraction. When extracting a given artistic style, we leverage ChatGPT (cha) to generate 10 common imagery. During the unlearning phase, at each epoch, we select one of these imagery to generate 8 images through prompts such as " The painting of [imagery] by [artist]". Similarly, during memorization, we select imagery to generate 8 images with prompts like " The painting of [imagery]" while including negative prompts like "by [artist]". For the *extraction* of a particular IP character, our approach involves generating 8 images through prompts like "The cartoon of the [IP character]" for the unlearning process. Similarly, during memorization, we utilize prompts such as "The cartoon of the [character]" to generate 8 images.

Regardless of unlearning or memorization, our training regimen encompasses 10 epochs, with each epoch comprising 30 iterations. We use a learning rate of 1.5e-4, a timestep value of 20 for the diffusion process, and a rank of 40 for LoRA. In *combination* phase, we adopt a learning rate of 1e-3 and utilize a larger rank value of 140 for LoRA.

Metric. To evaluate the efficacy of the *extraction*, we measure the discrepancy between the image sets generated by the base model and that generated by the non-infringing model after extraction with the same set of prompts. We want the discrepancy large when the prompts are with target concepts and the discrepancy small with surrounding concepts. This means that the *extraction* operation accomplishes the goal: the non-infringing model cannot generate images with target concepts but can still generate high-quality images with other surrounding prompts.

We note that for the image generation task, the ultimate criteria is human evaluation and hence we present the generated images of various scenarios for readers. Nonetheless, to save the cost and to compare with existing approach, we adopt an objective metric, i.e., the *Kernel Inception Distance* (KID) (Bińkowski et al., 2018), to measure the above discrepancy, similar to the *Fréchet Inception Distance* (FID) (Heusel et al., 2017) but with arguably less bias and asymptotical normality. Moreover, we employ the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) to quantify the disparity of artistic style artworks. LPIPS is a robust measurement tool that effectively captures differences in human perception between two images, offering a more comprehensive evaluation of stylistic variations in generated artworks.

Baseline. We compare our method with the concept ablation approach (Kumari et al., 2023) and Erased Stable Diffusion (ESD) (Gandikota et al., 2023), which achieve good concept removal performance by aligning the latent representation of target concepts with that of anchor concepts. Algorithmically, they tune the whole model to remove target concepts rather than the LoRA tuning in our paper.

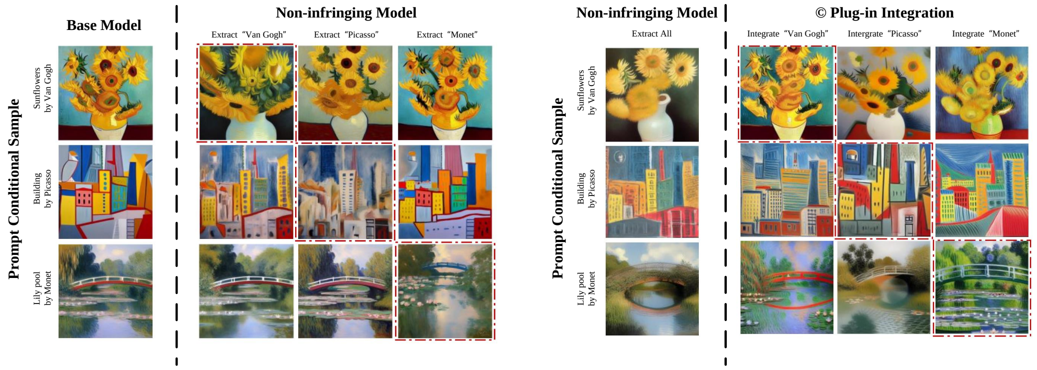
In general, we find it hard to compare the results with existing methods because of the complex setups in image generation, e.g., the tuning steps and the trade-off between removing the target concept and keeping the surrounding concept. Therefore we consider only the generation with similar scenarios and compare them under the same metric in the original paper.

D.2 Extraction and Combination of artist-style replication

Extraction. We extract artist styles from the Stable Diffusion V1.5, referred to as the "base model".Our focus revolves around three renowned artists: (1) Vincent van Gogh, (2) Pablo Ruiz Picasso, and (3) Oscar-Claude Monet. The outcomes of individual extraction are visually presented in Figure 7(a). The showcased images encompass those generated by both the base model and the non-infringing model, incorporating both the target style and surrounding styles.

In Figure 7, the images within red boxes represent the target styles, while the rest embody surrounding styles. A notable contrast is observable between the images within the red boxes and those generated by the base model. However, the images representing surrounding styles exhibit a substantial similarity to those generated by the base model. This illustrates the success of the *extraction* operation in isolating the target style from the base model while preserving the quality of images with surrounding styles.

In Table 2, we employ quantitative metrics to assess the efficacy of the *extraction* operation in contrast to the baseline methods. Our method showcases notable improvement, with the KID metric raising from 42 to 187 on the target style when compared to Concepts-Ablation (Kumari et al., 2023). This increase indicates enhanced removal of the target style. Additionally, in a comparative evaluation with the Erasing method (Gandikota et al., 2023), our method achieves a reduction in LPIPS from 0.21 to 0.14 on surrounding styles. This reduction implies less damage to the surround-



(a) Results of *extraction* in artist-style replication

(b) Results of *combination* in artist-style replication

Figure 7: **Results of style transfer**. In Figure (a), we show samples from different non-infringing models in each column. Each non-infringing model exhibits a deficiency in one style generation ability, with all other style generation capabilities remaining unaffected. In Figure (b), we present samples generated after adding certain ©Plug-ins in each column. Each of these ©Plug-ins serves to exclusively restore the generation of one particular style, while the generation of other styles continues to exhibit diminished performance.

Table 2: **Quantitative comparison with baselines in artist-style replication.** Compared to Concepts-Ablation, ours removes the target style more thoroughly, and compared to ESD, ours has less damage to surrounding styles.

Metrics	Methods	Target style ↑	Surrounding style ↓
KID$\times 10^3$	EXTRACTION (OURS)	**187**	32
	CONCEPTS-ABLATION	42	12
LPIPS	EXTRACTION (OURS)	0.31	**0.14**
	ESD	0.38	0.21

ing artistic styles, affirming our method's capability to safeguard the quality of the generated images with surrounding style prompts.

Combination. We combine the above three ©Plug-ins to construct a non-infringing model devoid of these three artistic styles associated with Van Gogh, Picasso, and Monet. Consequently, we individually integrate each style ©Plug-in to this model.

The non-infringing model in Figure 7(b) represents the model after extracting all of the styles of Van gogh, Picasso, and Monet. Notably, all the images generated by the non-infringing model markedly differ from those generated by the base model in Figure 7(a). This underscores the efficacy of the combination method, enabling the generative model to simultaneously exclude multiple styles.

The images enclosed in red boxes denote the target style after integrating the corresponding ©Plug-in, while the remaining images depict the surrounding style. The images of the target style are notably distinct from those generated by the non-infringing model and bear a closer resemblance to the images generated by base model in Figure 7(a). This observation shows that ©Plug-in can restore the model's capability to generate artworks of target style without contravening copyright restrictions associated with other artistic styles.

D.3 Experiment on ordinary objects generation

We evaluated the influence of *extraction* on the generation of ordinary objects. We utilize 5000 textual captions selected from the validation set in MS-COCO (Lin et al., 2015) as prompts, generating 5000 images using SD1.5 and the non-infringing model that extracts R2D2 and Picasso, respectively. Several randomly selected images are displayed in Figure 8. For illustrative purposes, we also generated results for concept-ablation and ESD on MS-COCO, respectively. Images within the same column exhibit substantial similarity, indicating that extraction does not exert an impact on the generation of ordinary items.

Table 3: **Quantitative results on MS-COCO.** FID and KID metrics for removing the Picasso style are presented in the upper two rows, while those for removing R2D2 are displayed in the lower two rows.

Domain	**Method**	**FID** $\downarrow$	**KID**$\times 10^3 \downarrow$
Style replication	*Extract Picasso*	24.04	2.83
	Erase Picasso	25.20	3.39
IP recreation	*Extract R2D2*	20.55	2.36
	Ablate R2D2	18.97	1.34

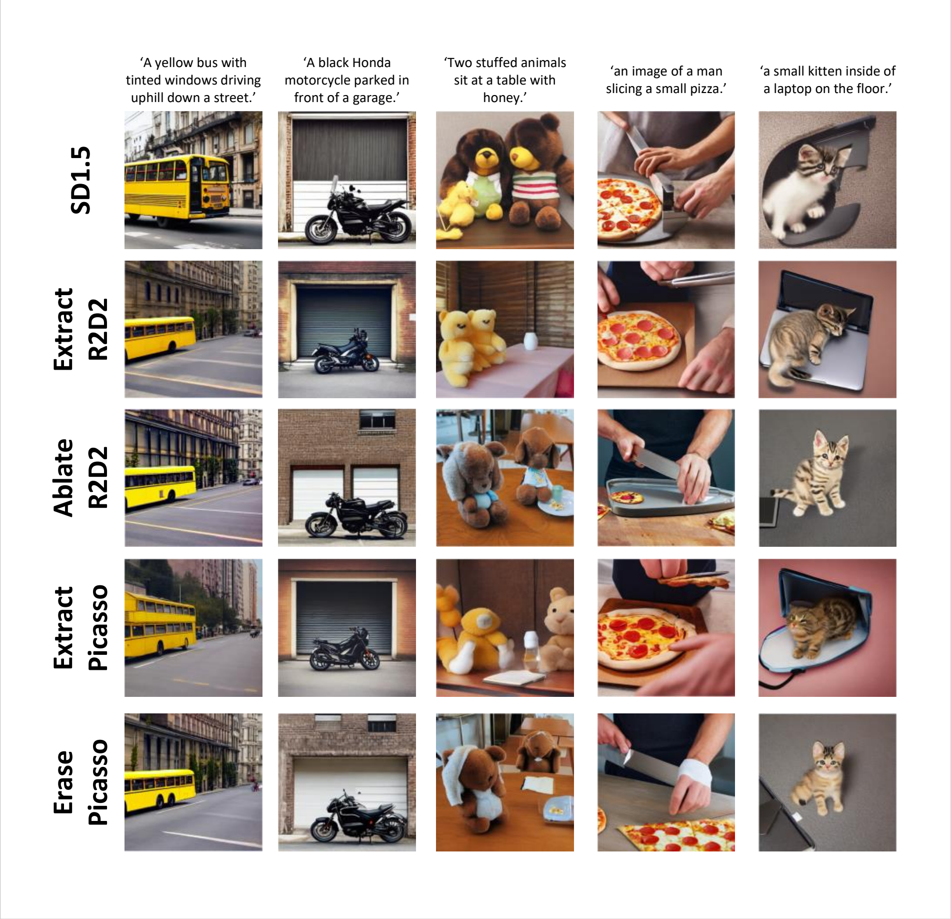

'A yellow bus with tinted windows driving uphill down a street.'
'A black Honda motorcycle parked in front of a garage.'
'Two stuffed animals sit at a table with honey.'
'an image of a man slicing a small pizza.'
'a small kitten inside of a laptop on the floor.'
SD1.5
Extract R2D2
Ablate R2D2
Extract Picasso
Erase Picasso

Figure 8: **Ordinary objects generation after *extraction*.** Row 1 displays images generated by Stable Diffusion V-1.5. Rows 2 and 3 illustrate images generated after the removal of the IP character R2D2, while Rows 4 and 5 showcase images generated after the elimination of Picasso's style. Rows 3 and 5 serve as the baseline, representing concept-ablation and ESD, respectively. Notably, after the *extraction* of R2D2 and Picasso, the non-infringing model retains the capability to generate commonplace objects sourced from the MS-COCO dataset (Lin et al., 2015).

As depicted in Table3, we calculate some quantitative metrics like FID and KID. It is noteworthy that Concept-Ablation (Kumari et al., 2023) only releases the checkpoint of ablating "R2D2" and ESD (Gandikota et al., 2023) only releases the checkpoint of erasing "Picasso". Thus, we compare them separately by using their respective checkpoints. Evaluation metrics consistently maintain low values, further affirming that extraction does not compromise the generation of ordinary items.