

# WOUAF: Weight Modulation for User Attribution and Fingerprinting in Text-to-Image Diffusion Models

## Supplementary Material

### A. Additional Related Work

**Text-to-Image Generative Models** Recent advancements in vector quantization and diffusion modeling have significantly enhanced text-to-image (T2I) generation, enabling the creation of hyper-realistic images from textual prompts [16, 20–22]. These T2I models have been effectively utilized in various tasks such as generating images driven by subject, segmentation, and depth cues [2, 3, 6, 11, 18]. However, the substantial size of these models presents a challenge for broader user adoption. Research efforts are focusing on enhancing model efficiency through knowledge distillation, step distillation, architectural optimization, and refining text-to-image priors [12, 15, 19, 23]. Amidst these technological advancements, ensuring the responsible usage of these powerful tools is a critical area of focus, which is the aim of our proposed method.

**Image Watermarking** Image watermarking aims to embed a watermark into images for asserting copyright ownership. To maintain the original image’s fidelity, these watermarks are embedded imperceptibly. Traditional approaches often employ Fourier or Wavelet transforms, while recent advancements leverage deep neural network-based auto-encoders for this purpose [26, 29, 30]. However, as discussed in the main paper, these methods can be easily disabled in an open-source setting.

From the standpoint of ownership verification, the fingerprinting of generative models aligns conceptually with watermarking techniques. However, unlike direct image manipulation to embed an identifiable signal in watermarking, generative model fingerprinting embeds this signal within the model’s weights. Consequently, the identifiable signal is integrated during the image generation process, akin to leaving fingerprints. This approach inherently prevents users from dissociating the fingerprinting process from image generation.

**Neural Network Watermarking** Watermarking techniques, particularly those embedding unique identifiers within model parameters, have been substantively explored in various studies, such as those highlighted in [1, 4, 17, 27, 28]. Our methodology, while aligning with the foundational principles of these works, introduces notable advancements in several key areas: utility, scalability, and verification methodology. The majority of existing watermarking techniques are tailored towards image classification models,

with only a limited subset extending their applicability to generative models, each presenting its own set of limitations. Unlike traditional methods that predominantly target single classification models, our approach endeavors to fingerprint approximately 4 billion Text-to-Image generator instances through a singular fine-tuning process. Additionally, while prior works have embedded fingerprints into various model aspects, such as input-output dynamics [1, 17] or directly within model weights [4, 27, 28], our strategy diverges by eliminating the necessity for trigger input, thereby enhancing scalability. In the context of our problem domain, where malicious users rarely share their model weights with the distributor responsible for watermark verification, the distributor typically only has access to potentially misused images. In essence, our approach not only aligns with but also extends beyond the conventional boundaries of network watermarking techniques, ensuring a thorough inclusion and discussion of these foundational methods in our related works section.

### B. Additional Details

WOUAF is evaluated utilizing the Stable Diffusion (SD) model [21] (version 2-base), trained specifically for generating images of 512p resolution.

### C. Additional Experimental Results

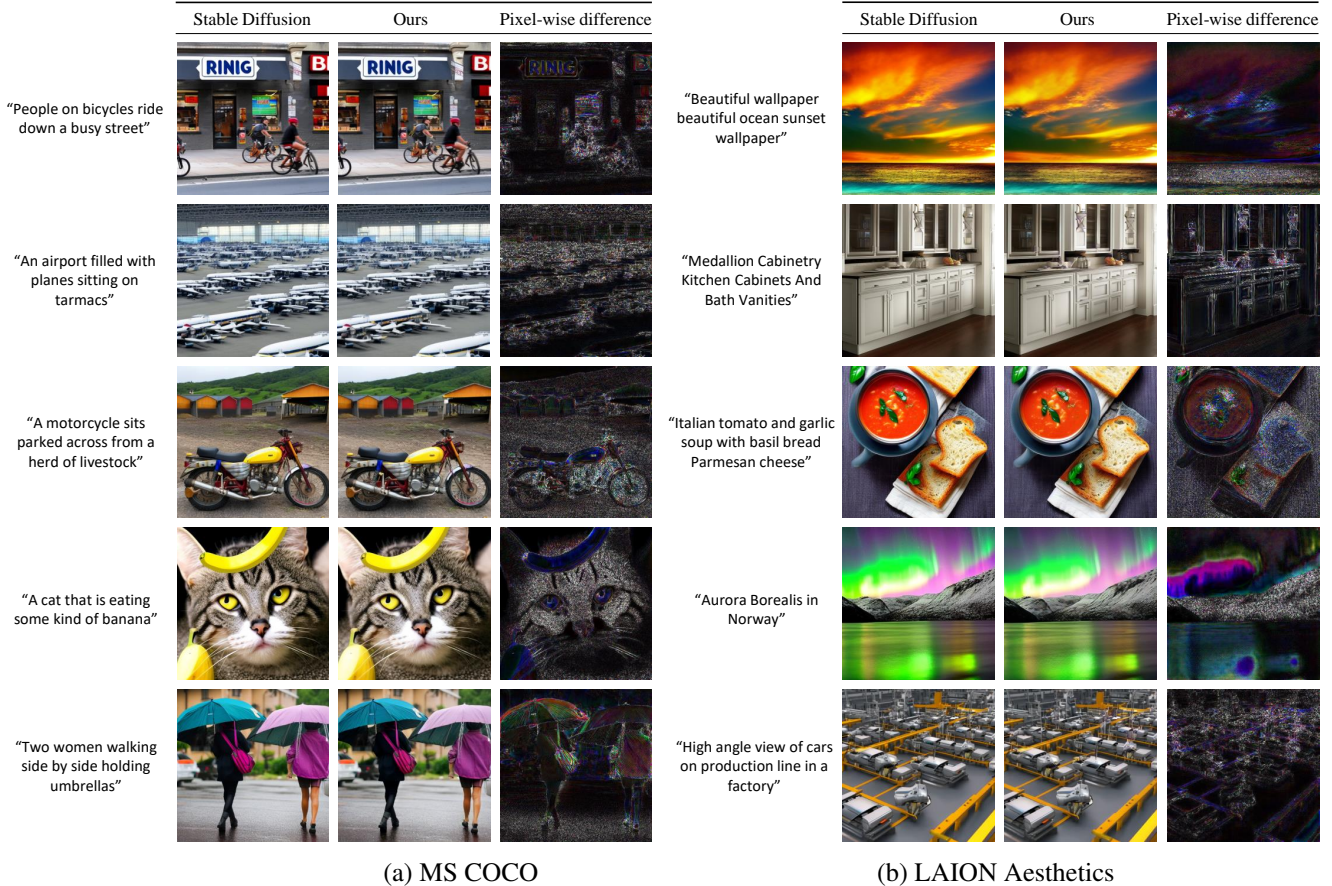
In addition to the figure in the main paper, we added uncurated images using text-prompt from MS-COCO [13] and LAION Aesthetics [24]. For convenience, we have aligned the subsection names with those in the main manuscript. Unless otherwise specified, all figures were generated using the ‘WOUAF-all’ method.

#### C.1. Additional Training Details

The dimension of the mapping network  $d_M$  is set to be equal to  $4 * d_\phi$  across all experimental setups. Training is performed over 50K iterations with a batch size of 32 and a learning rate of  $10^{-4}$  using AdamW optimizer [14].

#### C.2. Attribution Accuracy and Image Quality

As highlighted in the main manuscript, our methodology has a negligible effect on the original Stable Diffusion’s image quality. Please refer to Fig. 1 for these uncurated images.



(a) MS COCO

(b) LAION Aesthetics

Figure 1. Uncurated images of the original and fingerprinted Stable Diffusion models on MS-COCO and LAION Aesthetics. Pixel-wise differences are multiplied by a factor of 5 for a better view.

### C.3. Evaluating Generalizability Across Datasets

A key feature of our proposed methodology is its design independence from image-text paired datasets for achieving attribution accuracy. This property imbues it with the potential for broad applicability across a diverse range of contexts. To substantiate this claim, we conducted an experiment in which our variant models were trained exclusively on the ImageNet dataset [5]. We subsequently evaluated the performance of these ImageNet-trained models on the MS-COCO test set as well as a randomly selected portion of the LAION-aesthetics datasets.

The evaluation results, as seen in Table 1, effectively corroborate our assertion. Our methodology demonstrates compelling performance, with both our variants, achieving high attribution accuracy and maintaining image generation quality. These results underscore our method’s independence from the use of text-image paired datasets, thereby establishing its broad applicability in diverse scenarios where reliable fingerprinting and high-quality image generation are required. Fig. 2 provides a visual representa-

tion of these images.

### C.4. Attribution Accuracy Across Various Generation Hyperparameters

In accordance with the details provided in the primary manuscript, we subjected our methodology to evaluation employing two widely accepted schedulers: Euler [10], featuring time steps at intervals of [15, 20, 25], and DDIM [25], operating at time steps in [45, 50, 55]. Along with these, we also incorporated classifier-free guidance scales [9] at 2.5, 5.0, and 7.5.

Echoing the discussions in the main paper, the data in Tab. 2 and 3 corroborate the near-perfect attribution accuracy achieved by our method. Furthermore, the absence of significant deterioration in quality metrics reaffirms the resilience of our approach in the face of diverse generation hyperparameters (Refer to Fig. 3 and Fig. 4).

### C.5. Benefits of Finetuning only Decoder

In this section, we present qualitative outcomes resulting from the joint fine-tuning of the Stable Diffusion model’s

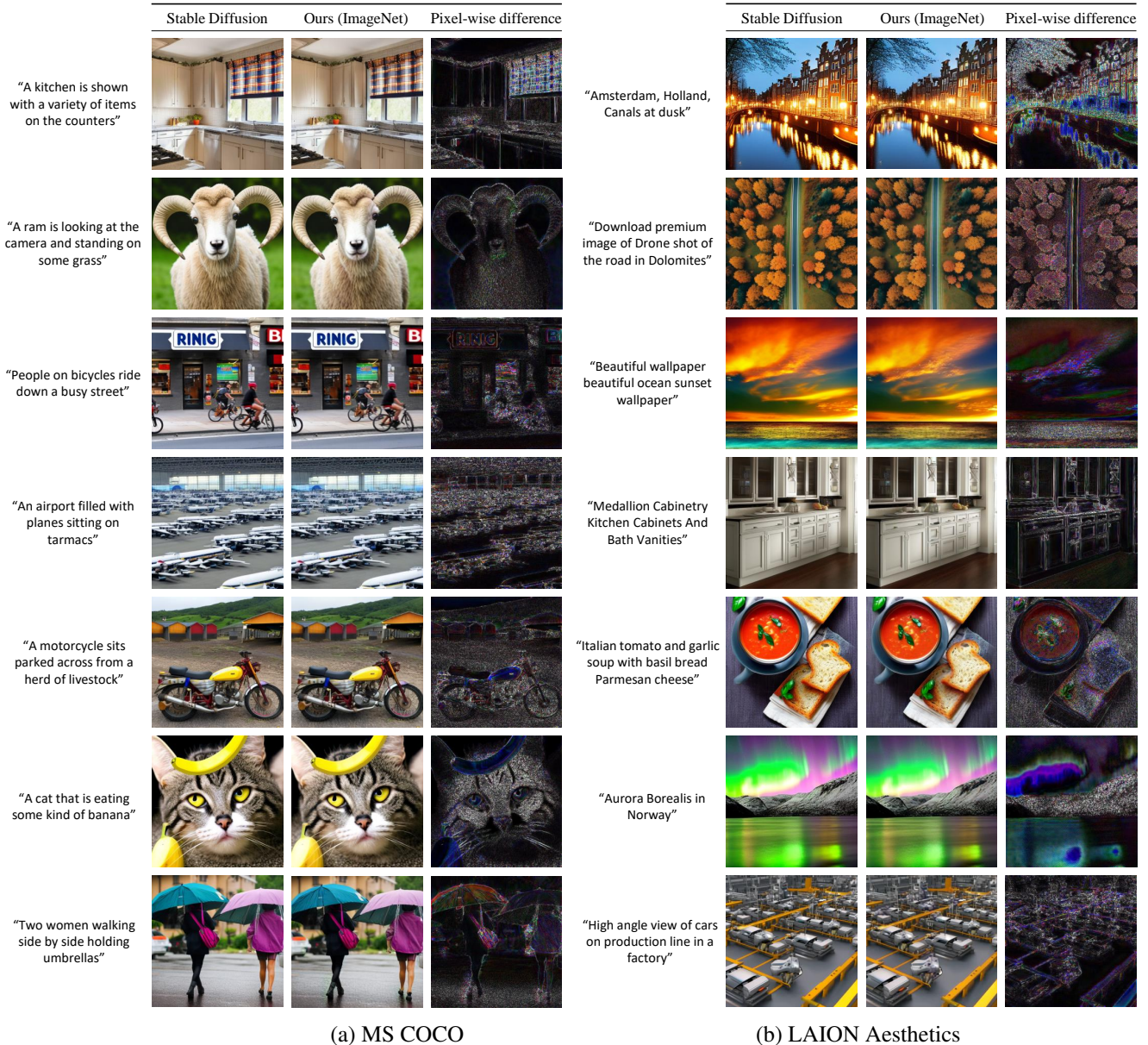


Figure 2. Qualitative comparisons of the original and fingerprinted Stable Diffusion models that were fine-tuned using only the ImageNet dataset. Pixel-wise differences are multiplied by a factor of 5 for a better view.

Table 1. Assessment of attribution accuracy and generation quality using Imagenet trained models. We validated our method using MS-COCO testset and LAION-aesthetics dataset.  $\uparrow/\downarrow$  indicates higher/lower is desired.

Model	MS-COCO			LAION		
	Attribution Acc. ( $\uparrow$ )	Clip-score ( $\uparrow$ )	FID ( $\downarrow$ )	Attribution Acc. ( $\uparrow$ )	Clip-score ( $\uparrow$ )	FID ( $\downarrow$ )
Ours-conv.	0.99	0.73	24.23	0.99	0.51	19.71
Ours-all	0.99	0.73	24.41	0.99	0.51	19.46

components, diffusion model  $\epsilon_\theta$  and decoder  $\mathcal{D}$ . As accentuated in the primary manuscript, our training proto-

col achieved an accuracy of 89%, however, it resulted in a noticeable deterioration in the quality metrics (Clip-score:

Table 2. Assessment of attribution accuracy and generation quality using Euler and DDIM scheduler with different time steps on MS-COCO. We fixed classifier-free guidance scale [9] to **7.5**.  $\uparrow/\downarrow$  indicates higher/lower is desired.

Model	Euler [10]				DDIM [25]			
	Steps	Attribution Acc. ( $\uparrow$ )	CLIP-score ( $\uparrow$ )	FID ( $\downarrow$ )	Steps	Attribution Acc. ( $\uparrow$ )	CLIP-score ( $\uparrow$ )	FID ( $\downarrow$ )
Original SD [21]	20	-	0.73	24.48	50	-	0.73	23.33
WOUAF-conv	15	0.99	0.73	24.63	45	0.99	0.73	23.28
	20	0.99	0.73	24.43	50	0.99	0.73	23.35
	25	0.99	0.74	24.14	55	0.99	0.73	23.31
WOUAF-all	15	0.99	0.73	24.65	45	0.99	0.73	23.34
	20	0.99	0.73	24.42	50	0.99	0.73	23.29
	25	0.99	0.73	24.11	55	0.99	0.73	23.26

Table 3. Assessment of attribution accuracy and generation quality on different classifier-free guidance scales [9] using MS-COCO. We fixed the scheduler and time steps to Euler for **20** steps and DDIM for **50** steps.  $\uparrow/\downarrow$  indicates higher/lower is desired.

Model	Guidance Scale 2.5				Guidance Scale 5.0			
	Scheduler	Attribution Acc. ( $\uparrow$ )	CLIP-score ( $\uparrow$ )	FID ( $\downarrow$ )	Scheduler	Attribution Acc. ( $\uparrow$ )	CLIP-score ( $\uparrow$ )	FID ( $\downarrow$ )
WOUAF-conv	Euler	0.99	0.72	18.63	Euler	0.99	0.73	21.91
	DDIM	0.99	0.72	18.35	DDIM	0.99	0.73	20.78
WOUAF-all	Euler	0.99	0.71	18.64	Euler	0.99	0.73	21.89
	DDIM	0.99	0.71	18.31	DDIM	0.99	0.73	20.68

0.68, FID: 63.48). Fig. 5 provides additional visual affirmation of these quantitative results.

### C.6. Robust User Attribution against Image Post-processes

We conducted a thorough evaluation of quality metrics to assess the impact of our robust user attribution training on various image post-processing methods. Examples of images post-processed using these methods are displayed in Fig. 6. As indicated in Tab. 4 and Tab. 5, our robust fine-tuning approach generally preserves image quality with only minimal perturbations. A representative example under a JPEG attack, generated by our robust model, is showcased in Fig. 7. Additionally, our method demonstrates adaptability under Combination attacks, which significantly challenge image fidelity. As illustrated in Fig. 6, these combined post-processing techniques necessitate a relatively stronger fingerprint compared to single post-processes, as further detailed in Fig. 8. Moreover, it is observed that images subjected to extensive post-processing lose perceptual value, impacting both malicious and naive users alike.

## D. Additional Deliberate Fingerprint Manipulation

### D.1. Gaussian Noise Model Purification

This subsection addresses the scenario where an adversary, upon recognizing the presence of fingerprints within the images generated by the image decoder  $\mathcal{D}$ , opts to add Gaussian noise into  $\mathcal{D}$  to obliterate the embedded fingerprint. In order to test this scenario, we gradually increase the standard deviation following  $[0., 0.01, 0.015, 0.02, 0.025, 0.03]$ . As shown in Fig. 10, our empirical analysis reveals a significant challenge: efforts to decrease the attribution accuracy lead to a decline in the quality of the generated images. This result also supports the idea that efforts to decrease attribution accuracy lead to a significant decline in the quality of the generated images.

### D.2. Full Knowledge Attack Scenario

This scenario assumes an internal attacker with comprehensive knowledge of our training process, including the training dataset, model structure, fingerprint space, and training details. To validate this, we trained an attacker’s version, following our methodology but employing a different random seed. We then assessed user attribution accuracy by inputting 5K images generated by the attacker’s model into WOUAF-conv and WOUAF-all fingerprint decoding net-

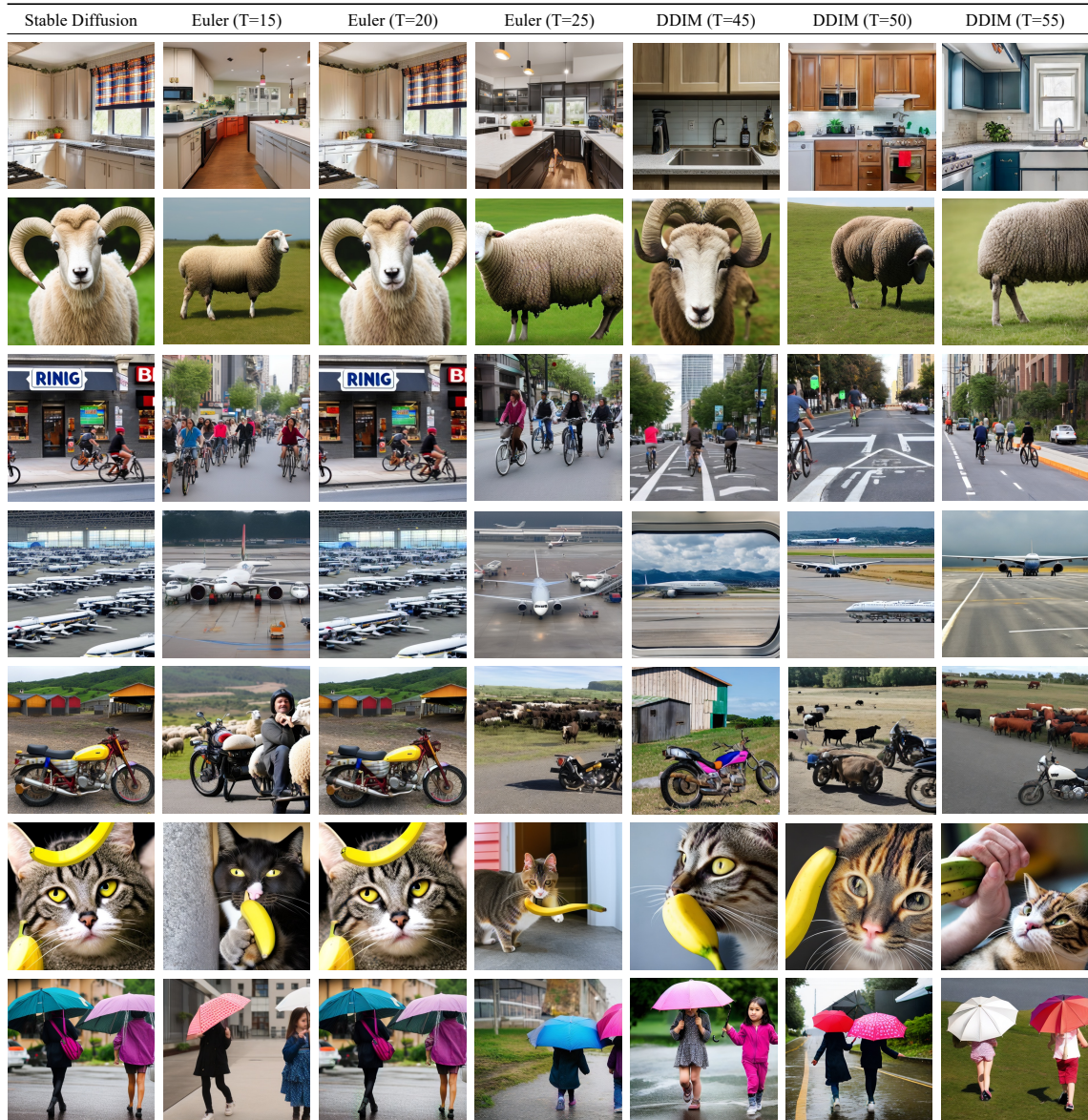


Figure 3. Qualitative results obtained using the Euler and DDIM schedulers with varying time steps on the MS-COCO dataset. We maintained a constant classifier-free guidance scale [9] at 7.5. Each column corresponds to the 'WOUAF-all' rows in Table 2.

Table 4. FID [8] scores using MS-COCO after robust training. Lower is desired.

Model	Crop	Rotation	Blur	Brightness	Noise	Erasing	JPEG	Combi.
WOUAF-conv (robust)	24.02	24.05	23.80	24.14	23.96	24.16	24.42	26.80
WOUAF-all (robust)	24.35	23.92	24.18	24.54	24.24	24.48	24.41	26.85

works. Both of our model variants exhibited user attribution accuracies of **0.509** and **0.501**, which are essentially random guesses, and thus dodged the attack. Even when an attacker with complete knowledge replicates our methodology, they

will not be able to mislead the original fingerprint decoding network.

Table 5. CLIP scores [7] using MS-COCO after robust training. Higher is desired.

Model	Crop	Rotation	Blur	Brightness	Noise	Erasing	JPEG	Combi.
WOUAF-conv (robust)	0.716	0.717	0.716	0.716	0.712	0.717	0.714	0.702
WOUAF-all (robust)	0.719	0.717	0.718	0.718	0.710	0.718	0.716	0.704














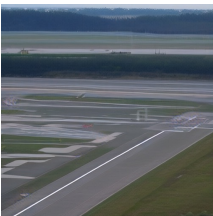






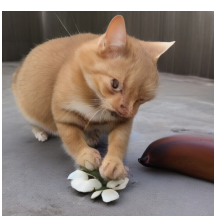

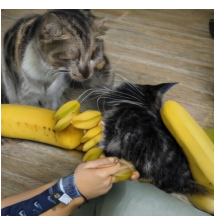
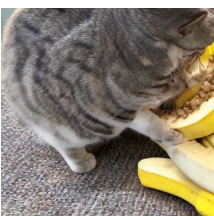
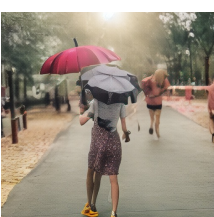



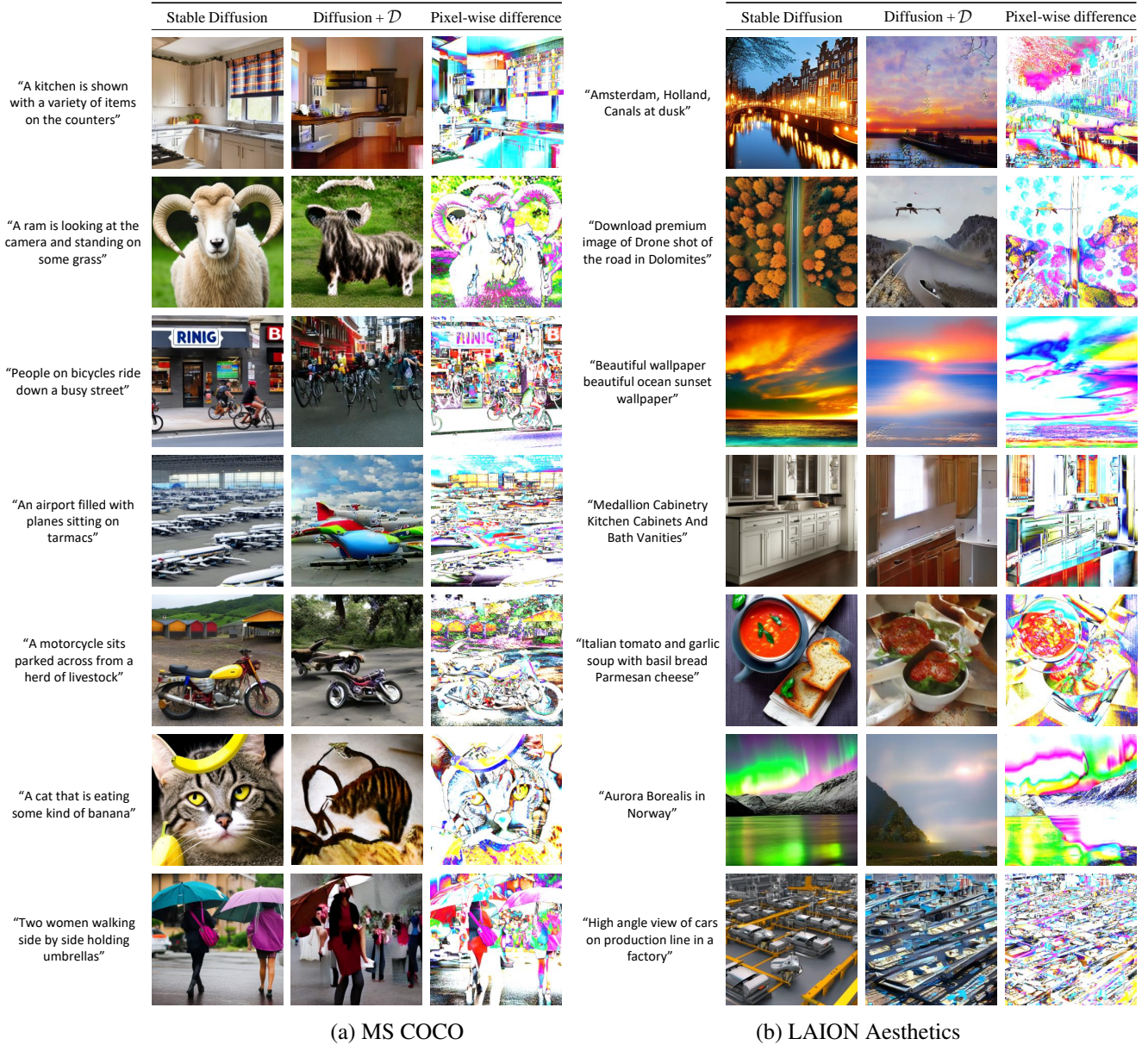
	Euler (s=2.5)	Euler (s=5.0)	DDIM (s=2.5)	DDIM (s=5.0)
“A kitchen is shown with a variety of items on the counters”				
“A ram is looking at the camera and standing on some grass”				
“People on bicycles ride down a busy street”				
“An airport filled with planes sitting on tarmacs”				
“A motorcycle sits parked across from a herd of livestock”				
“A cat that is eating some kind of banana”				
“Two women walking side by side holding umbrellas”				

Figure 4. Qualitative results produced by applying different classifier-free guidance scales [9] on the MS-COCO dataset. The scheduler and time steps were held constant at Euler for 20 steps and DDIM for 50 steps. Each column aligns with the ‘WOUAF-all’ rows in Table 3.



(a) MS COCO

(b) LAION Aesthetics

Figure 5. Qualitative results of the original and fingerprinted Stable Diffusion models on MS-COCO and LAION Aesthetics. When fine-tuning the SD model's  $\epsilon_\theta$  and  $\mathcal{D}$  together, there are significant quality drops. Pixel-wise differences are multiplied by a factor of 5 for a better view.



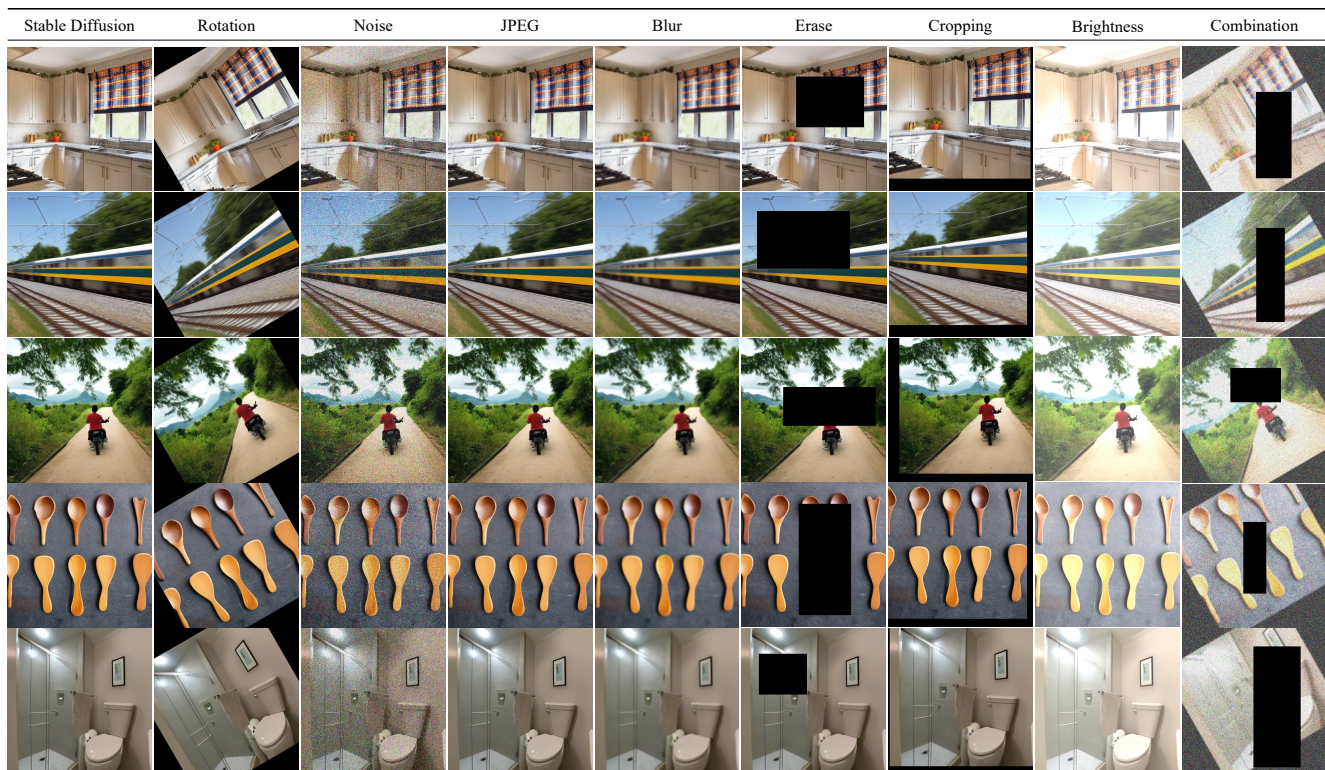
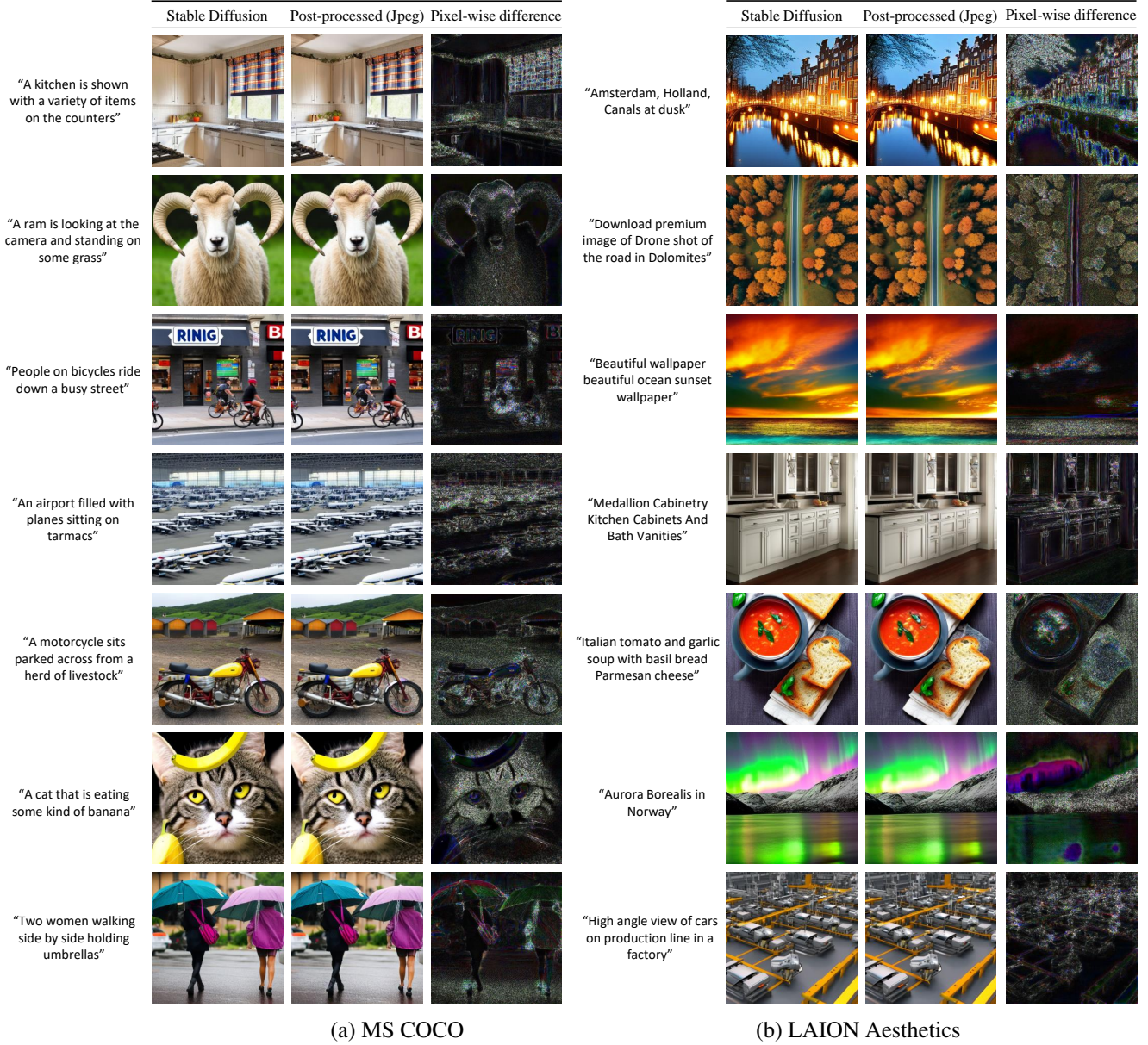


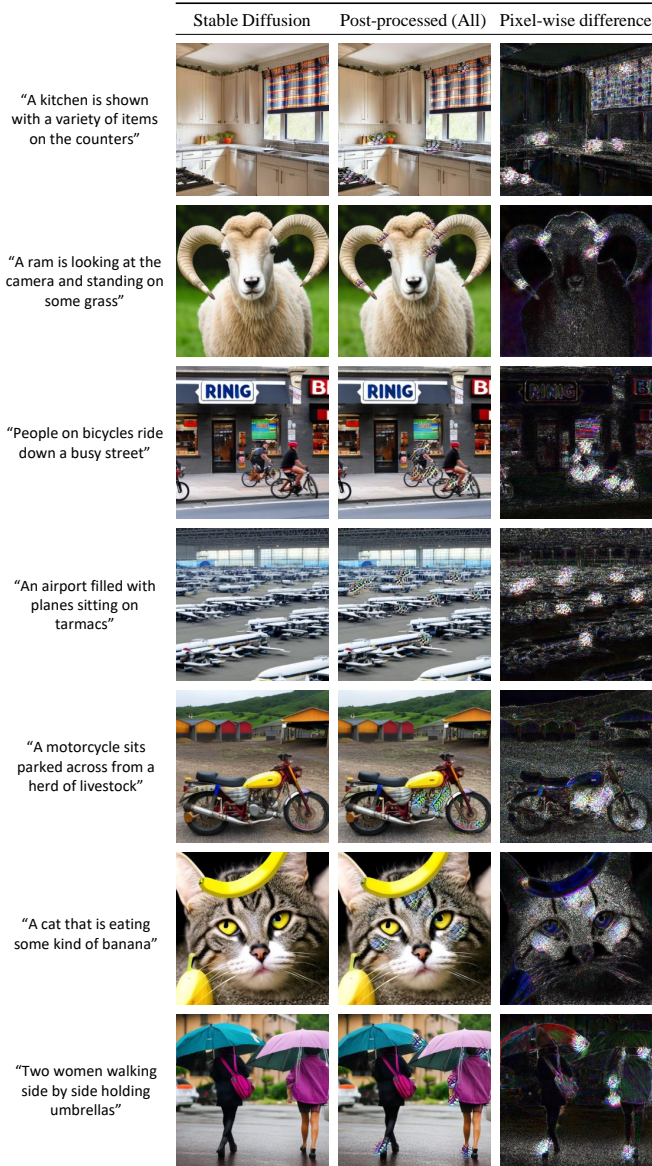
Figure 6



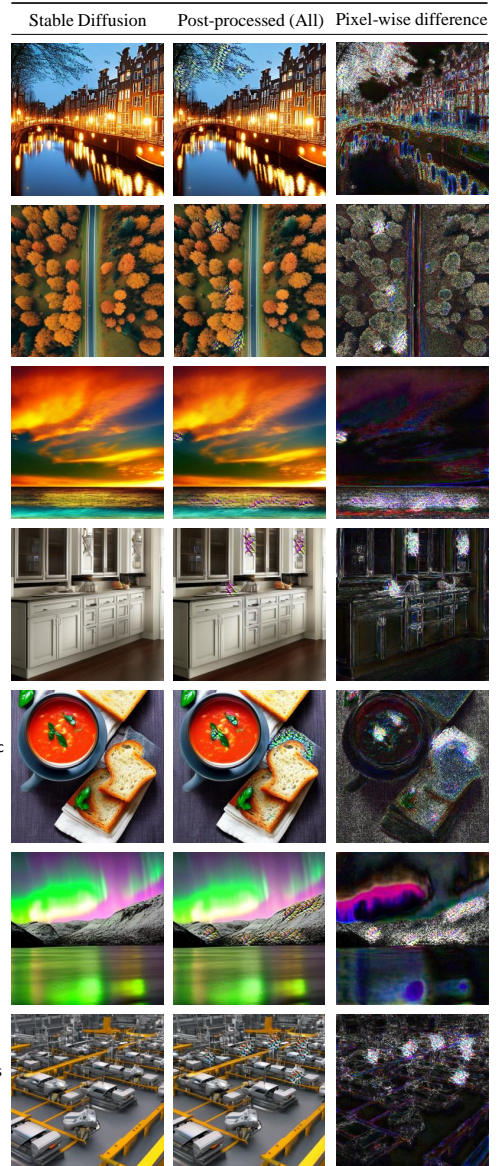
(a) MS COCO

(b) LAION Aesthetics

Figure 7. Qualitative results of the original and fingerprinted Stable Diffusion models on MS-COCO and LAION aesthetics. Our fingerprinted model is trained by simulating JPEG compression during training. Pixel-wise differences are multiplied by a factor of 5 for a better view.



(a) MS COCO



(b) LAION Aesthetics

Figure 8. Qualitative results of the original and fingerprinted Stable Diffusion models on MS-COCO and LAION aesthetics. Our fingerprinted model is trained by simulating all the combinations of the post-processing during training. Pixel-wise differences are multiplied by a factor of 5 for a better view.



Figure 9. Qualitative results of the original and fingerprinted Stable Diffusion models (WOUAF-conv) on MS-COCO. Pixel-wise differences are multiplied by a factor of 5 for a better view.

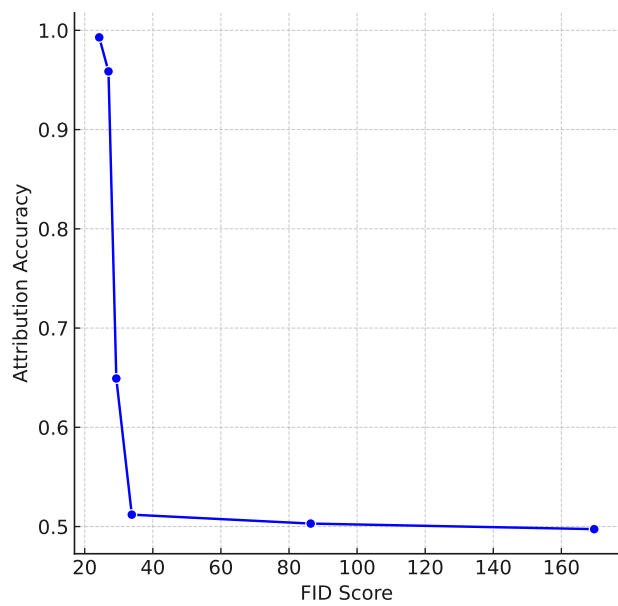


Figure 10. Model Purification. Adding Gaussian noise into weights leads to concurrent declines in both image quality and attribution accuracy. Note that a lower FID score is preferable, indicating better image quality.

## References

- [1] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 1615–1631, 2018. [1](#)
- [2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. [1](#)
- [3] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. [1](#)
- [4] Bitu Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 485–497, 2019. [1](#)
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [2](#)
- [6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. [1](#)
- [7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. [6](#)
- [8] 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. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. [5](#)
- [9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. [2](#), [4](#), [5](#), [7](#)
- [10] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. [2](#), [4](#)
- [11] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. [1](#)
- [12] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *arXiv preprint arXiv:2306.00980*, 2023. [1](#)
- [13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. [1](#)
- [14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [15] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. [1](#)
- [16] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. [1](#)
- [17] Ding Sheng Ong, Chee Seng Chan, Kam Woh Ng, Lixin Fan, and Qiang Yang. Protecting intellectual property of generative adversarial networks from ambiguity attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3630–3639, 2021. [1](#)
- [18] Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. *arXiv preprint arXiv:2306.04695*, 2023. [1](#)
- [19] Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: a resource-efficient text-to-image prior for image generations. In *ArXiv* –, 2023. [1](#)
- [20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. [1](#)
- [21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. [1](#), [4](#)
- [22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. [1](#)
- [23] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021. [1](#)
- [24] 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. *arXiv preprint arXiv:2210.08402*, 2022. [1](#)
- [25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [2](#), [4](#)
- [26] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2126, 2020. [1](#)

- [27] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 269–277, 2017. [1](#)
- [28] Tianhao Wang and Florian Kerschbaum. Riga: Covert and robust white-box watermarking of deep neural networks. In *Proceedings of the Web Conference 2021*, pages 993–1004, 2021. [1](#)
- [29] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. 2019. [1](#)
- [30] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 657–672, 2018. [1](#)