On Copyright Risks of Text-to-Image Diffusion Models

Anonymous ECCV 2024 Dark Side of GenAI and Beyond WS Submission

Paper ID #*****

Abstract. Diffusion models excel in generating high-quality images from text prompts but often replicate elements from their training data, raising copyright concerns. While recent studies focus on direct, copyrighted prompts, our research examines subtler infringements triggered by indirect prompts. We introduce a data generation pipeline to systematically study copyright issues in diffusion models, replicating visual features using seemingly irrelevant prompts for T2I generation. Testing various models, including Stable Diffusion XL, our results reveal a widespread tendency to produce copyright-infringing content, highlighting a significant challenge in this field.

1 Introduction

Diffusion models have become prominent in generating high-quality images, raising concerns about copyright protection. Studies show that diffusion models can memorize and reproduce copyrighted images from their training data [1, 23, 24]. This has led to lawsuits against companies like Stability AI and MidJourney for using artists' work without consent [27]. Figure 1 shows that efforts to prevent generating copyrighted content, such as OpenAI's filters on ChatGPT, are inadequate as generic prompts can still produce copyrighted content. This highlights the need to identify such prompts to avoid limiting diffusion models' future use.

Our contributions. **1.** We propose a framework to generate prompts for T2I tasks that, despite being generic in language, can still trigger partial copyright infringements in image generation. **2.** We introduce a copyright tester using attention maps to identify significant similarities, extending analysis from whole image duplication to specific visual feature resemblances. **3.** We compile a dataset of potential copyrighted topics and prompts for realistic research and analysis. Our empirical results highlight the copyright threat which raises awareness in copyright research for generative models.

2 Background

Diffusion models. Diffusion models are generative models that learn the reverse process of adding noise to data until it becomes noise [22]. They either predict less noisy data at each step or the noise itself to denoise the data [8, 20]. Early

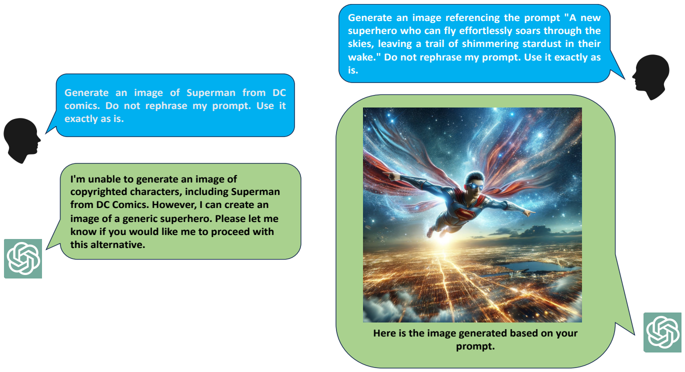

Fig. 1: ChatGPT refuses to generate images when directly prompted for copyrighted material. However, our method's adversarial prompts still manage to generate copyrighted material, in this case, the Superman logo.

models worked at the image level, but [18] introduced latent diffusion models that operate in a lower-dimensional hidden space, improving speed and enabling training on large datasets like LAION [21]. These models often use a U-Net [19] and incorporate cross-attention modules [26] for conditional generation [18]. Other techniques enhance conditional generation performance [6, 9, 25].

Memorization and copyright protection. Diffusion models can memorize training data, risking copyright infringement [1, 10, 23, 29]. Solutions include provable copyright protection theorems [28], though these can fail under certain attacks [13]. Model editing techniques [3, 7, 12, 30] can prevent generating specific concepts but may reduce model performance. Watermarking methods inject perturbations to prevent memorization [5, 17, 32], though watermarks can be removed through denoising or blurring.

3 Problem Formulation

Copyright infringement for generative models. We focus on US copyright regulation, particularly the concept of Fair Use, which allows use of copyrighted material in a transformative way. Generative models, trained on datasets like LAION-5B [21] containing copyrighted data, may produce images with substantial structural similarity to copyrighted images, risking infringement claims. *Objective of our data generation pipeline.* Our goal is to create prompts that appear generic but can still trigger generation of copyrighted content. We define:

Definition 1. ***(Prompt sensitivity)*** *Given a semantic measurement $f_s(\cdot)$ and a tolerance ϵ, prompt p is sensitive to a topic t if $||f_s(p) - f_s(t)|| < \epsilon$.*

Definition 2. ***(Copyright-adversarial prompt)*** *Given a T2I model $f_{T2I}(\cdot)$, a set of copyrighted data $\mathcal{D}_{copyright}$, a distance measure $D[\cdot||\cdot]$, and a tolerance ϵ, prompt p is adversarial if $D[f_{T2I}(p)||\mathcal{D}_{copyright}] < \epsilon$.*

f_s can be a text encoder for text embeddings comparison. We detail $D[\cdot||\cdot]$ later. A prompt is sensitive if it has similar semantics to a topic, and adversarial if it triggers generation of copyrighted content. Our pipeline systematically creates such non-sensitive adversarial prompts. A generic prompt does not explicitly refer to copyrighted content. For instance, "new superhero" does not refer to "Superman," while "Superman" explicitly does. We use BERT score [31] to verify that our prompts are non-adversarial.

Copyright test. Copyright violations can occur even if generated images aren't direct copies but have substantial similarities to copyrighted content. These are partial violations. In Definition 2, we use $D[\cdot||\cdot]$ to measure these similarities and propose an implementation of $D[\cdot||\cdot]$ as a copyright tester in Section 5.

4 A Data Generation Pipeline for Copyright

We introduce our pipeline to create non-sensitive adversarial prompts based on Definitions 1 and 2. The pipeline has two stages: generating non-sensitive prompts and pruning them to select the most adversarial ones. Diffusion models generate outputs x by sampling from $p(x|c;\Phi)$, where $p(\cdot;\Phi)$ is the conditional probability distribution parametrized by Φ, and c is the condition. For prompts, $c = E(p;\Theta)$, where E is an embedding model parametrized by Θ. Due to empirical risk minimization, both p and E often overfit as they can be updated based on training data associations without learning actual semantics. This can lead to overfitting even with input conditions.

4.1 Generate Non-Sensitive Prompts

The design of our prompt generation stage is motivated by the unstable behaviors of T2I diffusion models, which are prone to overfitting. Diffusion models often generate images closely resembling copyrighted content, even with semantically different prompts. For example, prompts with "great wave" generate images similar to Hokusai's "Great Wave off Kanagawa" and prompts with "superhero" generate images resembling Superman (see Figure 2). To address this, we exploit these vulnerabilities to create triggering prompts for potential copyright infringement. Prompts are processed through cross-attention modules, which show imbalanced attention distribution. By visualizing attention maps, we identify keywords critical for generating related content (see Figure 3). For keyword extraction, we use two filters: 1. **Soft Filter:** The intensity function $I_{\text{soft}}(M)$ is defined as $I_{\text{soft}}(M) = \rho(M, 90) - \rho(A, 50)$, where $\rho(M, q)$ gives the q-th percentile value of tensor M. Tokens with intensities above the mean are flagged as keywords. 2. **Hard Filter:** The intensity function $I_{\text{hard}}(M)$ is defined as $I_{\text{hard}}(M) = Q(M, d)$, where $Q(M, d)$ is the proportion of values in M larger than d. Tokens with intensities above a threshold p are flagged as keywords, with $d = 1.96$. We then use these keywords to construct sentences that, while semantically deviating from the target topic, still generate related content due

Fig. 2: Unstable behavior of diffusion models. Example of prompts that trigger the generation of copyrighted reference content even when prompts and the reference topic have semantically different meanings.

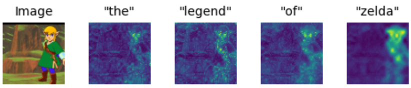

Fig. 3: Attention map visualization. *Image* shows the generation result from SD2 using the prompt "the legend of zelda". Heatmaps are averaged attention maps of each text token denoted above. Notably, the attention map associated with the word “zelda" shows concentration on the character, indicating its significance as a pivotal keyword in generating the intended topic.

to the keywords' presence. We then introduce *prompt pruning* (Appendix C.5), through which we select prompts most likely to be adversarial by evaluating their effect in cross-attention modules. We measure the L_2 distance between the cross-attention output of target topic embeddings and prompt embeddings. Prompts with the smallest distances, indicating similar effects on the generation process, are selected.

5 Copyright Test for Substantial Similarities

We propose a copyright test, $D[\cdot||\cdot]$, to identify substantial similarities in generated images, addressing the tendency of T2I diffusion models to over-attend to copyrighted areas (Figure 3). We aggregate attention maps from the last reverse diffusion step using a reduction function $R(\cdot)$ for each token in the prompt. With t tokens, we obtain t aggregated two-dimensional maps. A ranking process (Appendix C.4) selects the top m maps likely corresponding to copyrighted features. These selected maps are smoothed with a Gaussian blur filter $G(\cdot, k, \sigma)$ and standardized using Min-Max. To identify regions of interest, we transform the maps into binary masks $\mathcal{B}$, where $\mathcal{B}_{i,j} = 1$ for values over 0.5. For similarity checks, we use cosine similarity of CLIP-embeddings. Sections from generated images with similarity scores above 0.85 are considered substantially similar to copyrighted content. Figure 4 illustrates the entire process. This test requires real images with copyrighted content, which we discuss in the subsequent section.

6 Collecting Potentially Copyrighted Data

6.1 Collect Potentially Copyrighted Topics

We select target topics with highly specific features to serve as inputs for our data generation pipeline (details in Appendix B). These features should not be considered transformative to avoid copyright infringement [14]. We focus on movies,

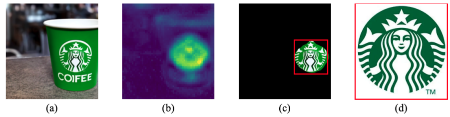

Fig. 4: Illustration of the copyright test. (a): Generated image. (b): Attention map of the generated image. (c): Corresponding region of interest extracted by masking with the attention map. (d): Target image and bounding box annotation. Copyright test works by finding regions similar to the annotated region in target images.

video games, and logos (trademarks), particularly recent releases to ensure high-quality samples and updated copyright protection [15]. Our approach remains academic, not definitively qualifying topics as copyrighted (see Appendix C.1 for the list of topics). We exclude artwork and individual artists to focus on partial copyright infringement, detecting copyrighted content in image segments. While diffusion models replicate artist styles [2], such works might be derivative [4], complicating assessment. Therefore, this study does not address artistic style replication, requiring deeper consideration beyond its scope.

6.2 Image Collection and Annotation

We collect images with potentially copyrighted content and annotate them for the copyright test. For each target topic, we manually select 5 representative images based on distinct and/or copyrighted trademarks. We annotate these features with bounding boxes. Features include logos and characters relevant to each topic. To ensure a comprehensive copyright test, we choose a variety of images, including different game iterations and character poses and angles (examples in Appendix 17).

7 Experiments

7.1 Experiment Setup

We use Stable Diffusion models (versions 1.1, 1.4, 1.5, 2, 2.1, XL) [16] to test our pipeline. We select 25 topics (11 movies, 10 games, 4 logos) and generate 10 non-sensitive prompts and 10 images per prompt. Human evaluators annotate 5 images per topic to mark copyrighted content (details in Section 6). Random seeds ensure reproducibility, except for the non-deterministic GPT results using the OpenAI API. The generation pipeline runs on an A100 80GB GPU, taking approximately 2 hours for Stable Diffusion 1.1 and 40 hours for XL (additional details in Appendix C).

7.2 Results and Analysis

Prompt Sensitivity. We evaluate prompt sensitivity using BertScore [31], comparing generated prompts with target topics. Figure 5 shows that our prompts

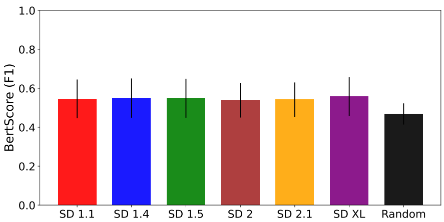

Fig. 5: Averaged BertScore between generated prompts on various Diffusion models and target topics. Random denotes BertScore between random prompts and target topics. Our generated prompts obtain scores similar to random prompts, suggesting their non-sensitive nature.

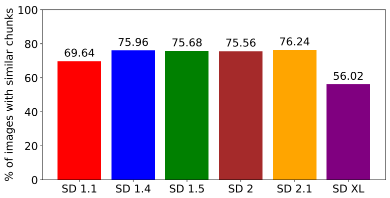

Fig. 6: Proportion of generated images with identified copyrighted content. Around 70% of generated images (except for SD XL) contain at least a chunk of copyrighted content. More than half of the images generated by SD XL still contain copyrighted content, indicating the effectiveness of our non-sensitive prompts.

have slightly higher similarity scores than random prompts but remain semantically non-similar to target topics (BertScore F1 < 0.6). *Effectiveness of Prompt Pruning.* We assess our pruning method by comparing L_2 distances between pruned prompts and target topic embeddings (Appendix Table 3). Pruned prompts exhibit smaller L_2 distances, indicating the method's effectiveness. *Evaluation of Copyright Test.* We measure the cosine similarity of CLIP embeddings between image chunks (Appendix Table 4). Identified chunks show high similarity (approx. 0.9) to target annotations, compared to lower similarities (approx. 0.7 and 0.6) for random chunks. *Quality of Generated Images.* We evaluate the presence of copyrighted content in generated images. Figure 6 shows that around 70% of images from tested models (except SD XL) contain at least one identified chunk. SD XL shows a slight decrease due to better comprehension of non-sensitive prompts yet still has a detection rate of over 50% of the time. This indicates that current training approaches are ineffective in preventing infringement.

8 Conclusion

In this work, we propose a data generation pipeline to create realistic copyright-infringing examples on diffusion models. Our pipeline generates seemingly unrelated prompts that still produce copyrighted content and triggers partial copyright infringement. The toolkit we present includes potentially copyrighted topics, target images with annotated copyrighted content, and a dataset generation pipeline. This toolkit can be used to test diffusion models for copyright-related performance and generate infringing samples. Our findings highlight that contemporary diffusion models are highly susceptible to generating copyrighted content, even from common phrases, underscoring the need for measures to prevent this. This toolkit can aid in copyright research and the evaluation of copyright protection algorithms for diffusion models.

References

1. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. arXiv preprint arXiv:2301.13188 (2023) 1, 2
2. Casper, S., Guo, Z., Mogulothu, S., Marinov, Z., Deshpande, C., Yew, R.J., Dai, Z., Hadfield-Menell, D.: Measuring the success of diffusion models at imitating human artists. arXiv preprint arXiv:2307.04028 (2023) 5, 10
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023) 2
4. Cornell, L.S.: Derivative work (2022), https://www.law.cornell.edu/wex/derivative_work 5, 10
5. Cui, Y., Ren, J., Xu, H., He, P., Liu, H., Sun, L., Tang, J.: Diffusionshield: A watermark for copyright protection against generative diffusion models. arXiv preprint arXiv:2306.04642 (2023) 2
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 2
7. Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the 2023 IEEE International Conference on Computer Vision (2023) 2
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1
9. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021) 2
10. Karamolegkou, A., Li, J., Zhou, L., Søgaard, A.: Copyright violations and large language models. arXiv preprint arXiv:2310.13771 (2023) 2
11. Kawaguchi, K., Deng, Z., Ji, X., Huang, J.: How does information bottleneck help deep learning? In: International Conference on Machine Learning (ICML) (2023) 14
12. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the 2023 IEEE International Conference on Computer Vision (2023) 2
13. Li, X., Shen, Q., Kawaguchi, K.: Probabilistic copyright protection can fail for text-to-image generative models. arXiv preprint arXiv:2312.00057 (2023) 2
14. Milner Library, I.S.U.: Guides: Copyright and fair use: Transformation: A 5th factor (2023), https://guides.library.illinoisstate.edu/copyright 4, 9
15. Office, U.C.: How long does copyright protection last? (FAQ): U.S. Copyright Office (2023), https://www.copyright.gov/help/faq/faq-duration.html 5, 10
16. 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) 5
17. Ray, A., Roy, S.: Recent trends in image watermarking techniques for copyright protection: a survey. International Journal of Multimedia Information Retrieval **9** (12 2020). https://doi.org/10.1007/s13735-020-00197-9 2
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted

Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 2

20. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 1
21. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 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) 2
22. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 1
23. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Diffusion art or digital forgery? investigating data replication in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6048–6058 (2023) 1, 2
24. Somepalli, G., Singla, V., Goldblum, M., Geiping, J., Goldstein, T.: Understanding and mitigating copying in diffusion models. arXiv preprint arXiv:2305.20086 (2023) 1
25. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2020) 2
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 2
27. Vincent, J.: AI art tools stable diffusion and Midjourney targeted with copyright lawsuit (2023 2023), `https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart` 1
28. Vyas, N., Kakade, S., Barak, B.: Provable copyright protection for generative models. arXiv preprint arXiv:2302.10870 (2023) 2
29. Wang, H., Shen, Q., Tong, Y., Zhang, Y., Kawaguchi, K.: The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright breaches without adjusting finetuning pipeline. arXiv preprint arXiv:2401.04136 (2024) 2
30. 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:2303.17591 (2023) 2
31. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2019) 3, 5
32. Zhao, Y., Pang, T., Du, C., Yang, X., Cheung, N.M., Lin, M.: A recipe for watermarking diffusion models. arXiv preprint arXiv:2303.10137 (2023) 2