Fantastic Copyrighted Beasts and Where to Find Them in Image-generation Models

Anonymous Authors[1]

Recent advances in image and video-generation models demonstrate remarkable ability for generating high-quality visual content based on free-form user inputs (Rombach et al., 2022; Betker et al., 2023; Chen et al., 2024; Li et al., 2024; Blattmann et al., 2023; Esser et al., 2024). However, they are susceptible to memorizing and generating entire datapoints or concepts from their training data (Somepalli et al., 2023; Carlini et al., 2023b;a). Since some training data originates from copyrighted materials (Carlini et al., 2023a; Kumari et al., 2023), regurgitation of such content may lead to legal intellectual property liability for users and model deployers who further make use of the generated content. In particular, this liability may stem not only from verbatim generation of training data, but generation of concepts highly similar to those from the training data, including copyrighted characters. Several legal scholars have highlighted that copyrighted characters will pose a particularly difficult, and under-explored, challenge for image or video generation services (Sag, 2023; Henderson et al., 2023). At least one lawsuit in China has already resulted in liability for an image generation service that generated the copyrighted character, Ultraman (Shimbun, 2024).

In this work, we provide a study focused on a subset of copyrighted content: *copyrighted characters*, such as popular IPs from Disney, Nintendo, and Dreamworks.[1] Given the legal risks involved and the need to respect individuals' intellectual property rights to copyrighted characters, commercial services like DALL·E have begun deploying interventions like prompt rewriting (OpenAI, 2024) to prevent generating copyrighted characters. However, such interventions have never been systematically assessed. Our work fills this gap by making the following key contributions.

First, we build COPYCAT , an evaluation suite with a diverse set of popular copyrighted characters and a novel evaluation pipeline, to better ground our investigation and evaluation. The suite contains (1) a dataset with 50 carefully-diverse popular copyrighted characters from 18 studios, both US and international, (2) GPT-4 based evaluator to identify copyrighted character in generated images, and (3) evaluation metrics capturing the inherent trade-off between copyright compliance and consistency with user intent in our evaluation: the DETECT score measures how many of 50 copyrighted characters a model can generate with specific inputs and interventions; the CONS score assesses if the main characteristic (e.g., "cartoon mouse" for Mickey Mouse) is present in the generation. We then evaluate five image generation models: Playground v2.5 (Li et al., 2024), Stable Diffusion XL (Podell et al., 2024), PixArt-α (Chen et al., 2024), DeepFloyd IF (StabilityAI, 2023), DALL·E 3 (Betker et al., 2023), and one video generation model, VideoFusion (Luo et al., 2023).

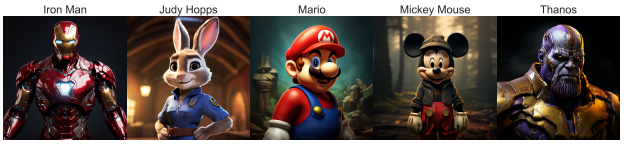
Figure 1. Examples of Mario generated by the open-source Playground v2.5 model and proprietary DALL·E 3 model. Mario can be generated with direct name inclusion (*character name anchoring*, not possible with DALL·E 3 due to guardrails) or using keywords (*indirect anchoring*, still possible with DALL·E 3).

Second, we demonstrate that as little as two generic keywords associated with a character are enough to generate their image, without mentioning their name. We formalize two modes of prompt to copyrighted character generation. The first type prompts with character name directly (*Character Name Anchoring*). The second type uses keywords or descriptions without the character's name (*Indirect Anchoring*). We include an example based on Mario in Figure 1. Indirect anchoring is especially important for both model deployers and model users: even a non-malicious user could accidentally generate copyrighted characters when using seemingly innocuous prompts, leading to potential legal liability for the model deployer as well as any unsuspecting user that tries to monetize the image.

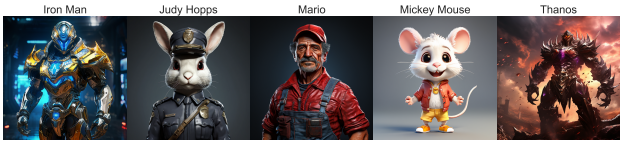
[1]We study the phenomenon of copyrighted characters generation from an image-based reasoning and natural language processing perspective. Artist compensation and other downstream policy issues are also important, but the formal discussion on these topics are beyond the scope of this work.

Negative Prompt	Prompt: Target's name DETECT (↓)	Prompt: Target's name CONS (↑)	Prompt: Rewritten prompt DETECT (↓)	Prompt: Rewritten prompt CONS (↑)
None	$30.33_{\pm 1.89}$	$0.75_{\pm 0.01}$	$14.33_{\pm 2.62}$	$0.80_{\pm 0.01}$
"Copyrighted character"	$30.33_{\pm 1.25}$	$0.74_{\pm 0.01}$	$17.33_{\pm 1.70}$	$0.80_{\pm 0.01}$
+ 5 LM-RANKED keywords	$30.33_{\pm 1.89}$	$0.71_{\pm 0.01}$	$14.33_{\pm 1.70}$	$0.80_{\pm 0.00}$
+ 5 EMBEDDINGSIM keywords	$28.00_{\pm 1.41}$	$0.72_{\pm 0.03}$	$15.67_{\pm 1.25}$	$0.80_{\pm 0.00}$
+ 5 CO-OCCURRENCE-LAION keywords	$27.33_{\pm 0.00}$	$0.73_{\pm 0.01}$	$14.33_{\pm 0.94}$	$0.80_{\pm 0.00}$
+ 5 CO-OCCURRENCE-LAION & 5 EMBEDDINGSIM keywords	$23.33_{\pm 3.30}$	$0.72_{\pm 0.03}$	$7.00_{\pm 1.63}$	$0.81_{\pm 0.00}$
Target's name	$23.67_{\pm 2.62}$	$0.76_{\pm 0.01}$	$7.67_{\pm 0.47}$	$0.81_{\pm 0.01}$
+ 5 LM-RANKED keywords	$25.00_{\pm 1.63}$	$0.74_{\pm 0.01}$	$7.00_{\pm 1.63}$	$0.81_{\pm 0.02}$
+ 5 EMBEDDINGSIM keywords	$22.67_{\pm 2.36}$	$0.73_{\pm 0.02}$	$5.67_{\pm 0.47}$	$0.80_{\pm 0.00}$
+ 5 CO-OCCURRENCE-LAION keywords	$20.67_{\pm 2.05}$	$0.75_{\pm 0.01}$	$5.00_{\pm 0.82}$	$0.81_{\pm 0.01}$
+ 5 CO-OCCURRENCE-LAION & 5 EMBEDDINGSIM keywords	$20.67_{\pm 0.47}$	$0.72_{\pm 0.03}$	$4.33_{\pm 0.47}$	$0.81_{\pm 0.00}$

Table 1. Performance of all intervention strategies on the Playground v2.5 model. We run each strategy three times, and report the mean and standard deviation of the number of detected copyrighted characters (DETECT, lower is better) and the consistency with user intent (CONS, higher is better). Including the character's name in the negative prompts is crucial for reducing DETECT. Combining prompt rewriting and negative prompts can effectively reduce DETECT from 30 to 5, without significantly degrading CONS.



(a) Prompt: Character's name, Negative Prompt: None



(b) Prompt: Rewritten, Negative Prompt: Character's Name & 5 EMBEDDINGSIM & 5 CO-OCCURRENCE-LAION keywords

Figure 2. Playground v2.5 generation without (a) and with (b) intervention. Prompt rewriting paired with negative prompts reduces the likelihood of generating images resembling copyrighted characters while preserving the user's intended key characteristics.

To identify indirect anchors, we first use a language model to generate candidate keywords and descriptions that might evoke the likeliness of a copyrighted character. Then, we use the following 3 reranking approaches to semi-automatically discover indirect anchors: (1) LM-RANKED: using greedy decoding to capture the inherent ranking of LMs. (2) EMBEDDINGSIM Ranking: rank by their embedding space distance to the copyrighted character's name. (3) CO-OCCURRENCE Ranking: rank by their co-occurrence with the character's name in popular training corpora.

We find that co-occurrence with the LAION dataset (Schuhmann et al., 2022) (CO-OCCURRENCE-LAION) is especially useful for identifying indirect anchors. Our findings suggest a mismatch between the level of generality in the prompt versus the specificity of the output: For example, 40% of the characters in our benchmark can be recreated with only 5 automatically selected keywords using our method, and some with as little as 2 keywords. These indirect anchors also extend to video generation models and product systems with built-in safeguards, like DALL·E.

Third, we find that existing mitigations are insufficient and suggest new strategies. In this paper we focus on runtime approaches only, assuming that models cannot be modified to remove copyrighted characters. We explore possible solutions model deployers can practically incorporate into a production system. We find that applying prompt-rewriting, an important technique for copyright protection adopted by model deployers, alone can only reduce the number of generated characters by half compared to no intervention. Many of the failed rewritten prompts also contain indirect anchoring keywords we identified. Instead, combining this strategy with negative prompting (i.e., steering models away from concepts like "red hat" during inference)—where a model proactively steers away from inferred character at generation time—significantly boost mitigation effectiveness. This combination reduces over 80% of characters compared to no mitigation (Figure 2). We also show in Figure 2 that a combination of mitigation strategies is able to strike a balance between enhancing copyright compliance and adhering to user intent.

We summarize the key takeaways for users and model deployers as follows:

- We call for more awareness of indirect anchoring, where models can generate copyrighted characters without explicitly mentioning the character's name. For deployers, this may evade safeguards relying on direct name detection. For users, these prompts can result in potential infringement and liability even when not intending to generate copyrighted characters.
- For model deployers who adopt mitigation strategies and intend to prevent the generation of copyrighted characters, we recommend investing in techniques beyond prompt rewriting. Our work suggests a relatively simple set of strategies, such as combining prompt rewriting and negative prompts, to make progress on this issue.

References

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra†, W., Dhariwal†, P., Chu, C., Jiao, Y., and Ramesh, A. Improving image generation with better captions. 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *USENIX Security*, 2023a.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *ICLR*, 2023b.

Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *ArXiv*, abs/2303.15715, 2023. URL https://api.semanticscholar.org/CorpusID:257771630.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023.

Li, D., Kamko, A., Akhgari, E., Sabet, A., Xu, L., and Doshi, S. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., and Tan, T. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023.

OpenAI. Image genexfration. https://platform.openai.com/docs/guides/images/usage, 2024.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Sag, M. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023.

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. In *NeurIPS*, 2022.

Shimbun, T. Y. China court awards damages over ai images resembling ultraman; service provider held liable for copyright infringement. 2024. URL https://japannews.yomiuri.co.jp/society/crime-courts/20240416-180611/#:~:text=The%20court%20ruled%20on%20Feb,halting%20generation%20of%20the%20images.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*, 2023.

StabilityAI. DeepFloyd IF. https://github.com/deep-floyd/IF, 2023.