Safer Prompts: Reducing IP Risk in Visual Generative AI

Lena Reißinger[a], **Yuanyuan Li**[b], **Anna-Carolina Haensch**[a] **and Neeraj Sarna**[b]

[a]Ludwig Maximilian University of Munich
[b]Munich Re

Abstract. Visual Generative AI models have demonstrated remarkable capability in generating high-quality images from simple inputs like text prompts. However, because these models are trained on images from diverse sources, they risk memorizing and reproducing specific content, raising concerns about intellectual property (IP) infringement. Recent advances in prompt engineering offer a cost-effective way to enhance generative AI performance. In this paper, we evaluate the effectiveness of prompt engineering techniques in mitigating IP infringement risks in image generation. Our findings show that Chain of Thought Prompting and Task Instruction Prompting significantly reduce the similarity between generated images and the training data of diffusion models, thereby lowering the risk of IP infringement.

1 Introduction

As generative AI (GenAI) becomes increasingly prevalent in real-world applications, concerns about its potential risks continue to grow. We focus on the risks associated with the replication of copyrighted material and restrict ourselves to image generation. The tangible legal and financial implications of these risks have sparked our current research. For instance, artists sued Stability AI (Midjourney and DevianArt) claiming that the companies' AI image generators produce images that are strikingly similar to their artworks [2]. In August 2024, a U.S. district judge (Orrick) ruled that the artists could proceed with the copyright infringement claims, underscoring the ongoing legal uncertainties surrounding AI-generated content [6].

These prevalent risks make model users resistant to fully exploiting the latest GenAI models. To promote a wide adoption of GenAI, risk management is crucial. Risk management has two aspects: a) risk quantification; and b) risk mitigation. Both aspects have already received attention in the literature.

Risk quantification: Risk quantification involves analysing the fraction of training images that are reproduced by the model. Since training images may be copy-righted, this leads to IP infringement risks for the end user. Consider for instance Stable Diffusion-1. When prompted using the captions of training images directly, about 2% of images it generates are highly similar to those in the training dataset [4, 24]; the results are very similar for the Stable Diffusion-2 model [25].

Risk mitigation: We focus on risk mitigation, which takes a step beyond risk quantification. The goal of risk mitigation is to reduce the probability of a model outputting copyrighted content. The risk could either be mitigated before deploying a model (on the model developer's side) or after (on the model's user side).

Pre-deployment strategies involve changing the data science process and are usually expensive. For instance, models could be trained on a de-duplicated data set, which usually reduces the risk during training time [25]. Note however that with this strategy test time risks still prevail. Model could also be made to unlearn copy righted data but that comes with a huge computational expense [31, 7, 8].

Post-deployment strategies are usually cheaper. Model users could be educated on using AI responsibly. Furthermore, before a user inputs a prompt, a "system message" could be added that aims to reduce the IP-infringement risks [13]. This is a kind of prompt engineering techniques [28, 5, 17] that have been popularly used to enhance GenAI model performance on diverse tasks. Another possibility is to check each generated output against a corpus of copyrighted images. Since the size of such a corpus is huge, this approach soon becomes infeasible.

We focus on prompt engineering and evaluate its effectiveness for IP risk mitigation—Figure 1 presents a snapshot of our results indicating that prompt engineering could substantially alter training data replication for vision models. Since GenAI outputs are sensitive to input prompts, we expect that via a carefully engineered prompt, IP infringement related risk can be reduced. Prompt engineering has already been explored for reducing hallucination in large language models [26, 3]. For vision models however, it is largely unexplored [10].

Prompt engineering for image generation and user accessibility has primarily developed through informal, trial-and-error approaches. In artistic communities, a common trend has emerged where prompts frequently take the form of "X in the style of Y," with Y referring to a particular artist or artistic movement. Based on this observation, [11] explored how modifications to this prompt pattern influence the behavior of the image generation model. Their experiments focused on the influence of different phrasings of the prompt, different random initializations, the number of iterations, and the style as well as the subject parameter.

In this paper, we want to contribute to ongoing research on prompting in visual diffusion models by applying various prompt engineering strategies for generating non-copyright-violating content and evaluating their performances. To the best of our knowledge, for IP risk mitigation involving vision models, ours is the first work of its kind that develops an extensive framework for exploring prompt engineering.

1.1 Current contributions

We make the following contributions. We consider four different prompt engineering strategies for vision models. For each of these strategies, via extensive numerical experiments, we study the extent to which IP-infringement risk could be reduced. While reducing this risk, since relevant and aesthetic outputs are preferable, we also study

the correlation of both of these attributes with risk mitigation. We conclude with actionable insights, which could help model users in responsibly using visual GenAI models.

Figure 1: Example of generated images using different prompting strategies. The first row shows Stable Diffusion 2's generations for prompts with the caption "Vincent Van Gogh Cafe Terrace At Night", ordered from left to right: *No Prompt Engineering* (baseline), *Task Instruction Prompting*, *Negative Prompting*, and *Chain of Thought Prompting*. The second row displays their closest matching images from the LAION-12M dataset. All generations using prompting engineering strategies have significant new elements comparing with their closest match in original dataset.

2 Background

This section presents the different prompting strategies we consider along with a similarity-based definition of IP infringement.

2.1 Prompting Strategies

A prompt is a set of instructions given to a GenAI model to tailor its behavior, enhance its functionality, or refine its outputs [28]. Prompt engineering refers to the process of designing these instructions effectively to optimize the model's performance. We consider the following four prompting strategies:

1. *Baseline/No Prompt Engineering:* here we directly use the captions of the training images to generate the outputs. The captions have little to no modification. We consider this to be a baseline strategy.
2. *Task instruction prompting:* This means adding a very detailed description of the task the model should perform in the prompts. For mitigating IP infringement risk, this includes steering the model towards creating novel elements to produce unique output, as well as avoiding the reproduction of recognizable content. The sample metaprompt from [13], formulates this explicitly as *"to avoid copyright infringements"*;
3. *Negative prompting:* This includes the concept of negation (no, not, nor) within the (baseline) hard prompt. The effect of this strategy on stable diffusion has already explored [9]. Notably, the authors illustrate an example of a prompt that requires the model to produce an image of a man without glasses. The prompt still resulted in an image of a man wearing glasses;
4. *Chain-of-thought prompting:* This enables the model with self-check mechanisms where a model evaluates its reasoning. This could potentially improve model's ability to generate unique, non-infringing images as outputs. This technique has already been used to reduce hallucinations in LLMs [12]. Furthermore, for

vision models, authors have already shown that using chain-of-thought with vision models can enhance fairness [21] and improve reasoning ability [30].

Table 1 provides an overview of our formulations under each prompt engineering strategy. The specific wording was refined through a trial-and-error process during our initial tests. While this method may not be entirely systematic, informal trial-and-error approaches, as described by [11], have so far been the primary way prompts for text-to-image models have been developed.

Table 1: Prompts for each strategies

Prompting Strategy	Prompt
No Prompt Engineering	Generate an image of $\{caption\}$.
Task Instruction Prompting	Create a visually distinctive, highly creative, and non-copyright-infringing depiction of $\{caption\}$. Focus on originality and incorporate entirely novel visual elements. Avoid using recognizable characters, logos, or copyrighted designs. Ensure the image is imaginative and unique.
Negative Prompting	Generate an imaginative and original image of $\{caption\}$. The image must not include realistic replication, no known art styles, no recognizable characters, and no copyrighted material.
Chain of Thought Prompting	1. Generate a creative and unique image of $\{caption\}$, focusing on originality and imaginative composition. 2. Incorporate completely novel elements into the image that are distinct from the training data and are unlikely to resemble any existing images. 3. Ensure every element in the image is visually distinct, creative, and does not replicate known styles, characters, or objects present in existing datasets. 4. Verify the final output aligns with the given caption while maintaining a high degree of creativity and uniqueness.

2.2 Detecting IP Infringement

Copyright issues in visual diffusion models arise when models replicate their training data within the image generation process. Therefore, to detect IP infringement, we compare the generated image with the images in the training data. Since training data is huge, a comparison to the entire dataset is prohibitively expensive. We therefore—similar to [25]—limit our comparison to a subset of the training data.

Image embeddings: To compare two images, we first encode them. Image embedding can happen on a content and style level. Some embeddings correspond rather to the object-level content of the generated images, whereas other perform better in capturing the artistic style of the generated images [19]. We focus on rather content based embedding because similarity based upon style might not be considered copyrighted [14].

To embed an image, we consider CLIP embedding [18]. CLIP embeddings are useful when dealing with text-to-image models because they map both images and text into a shared semantic space. This also allows for meaningful comparisons beyond pixel-level similarity. By using CLIP embeddings, we evaluate similarity not only at a visual level but also in terms of content and semantic meaning.

Similarity score: Once we have the image embedding, we compare images using a similarity score between the embeddings. Two images are considered to be similar when the similarity score between their embeddings is larger than certain threshold [4]. We con-

sider the cosine similarity between the CLIP embeddings as our similarity score, and a threshold of 0.85 as high IP infringement risk. Notice that [4] used the same similarity threshold to identify near duplicates in the training data for deduplication.

We briefly recall other possible choices for a similarity score: authors in [24] use split dot product of Self Supervised Copy Detection (SSCD) scores as similarity measure, where scores are predicted based on differential entropy regularization, see [16]. In [27], authors utilize the Euclidean $L2$ norm on pixel space and SSCD scores, while [4] employ cosine similarity between CLIP embeddings to identify near-duplicates.

3 Experimental Results

3.1 Design of Experiment

Model and dataset: We use Stability AI's Stable Diffusion 2 as an example. It has been widely studied in foundational research on training data replication in visual diffusion models, such as [25]. Stable Diffusion 2 was trained on the LAION-2B-en dataset, a subset of LAION 5B, which comprises approximately 2.3 billion image-text pairs primarily in english [20]. As mentioned earlier, pairwise comparing our output image with the entire training dataset would be infeasible. Therefore, we restrict ourselves to a subset of this training set referred to as LAION-Aesthetics 12M dataset [1]. This dataset was also used to fine-tune Stable Diffusion 2—see [23, 20] for further details.
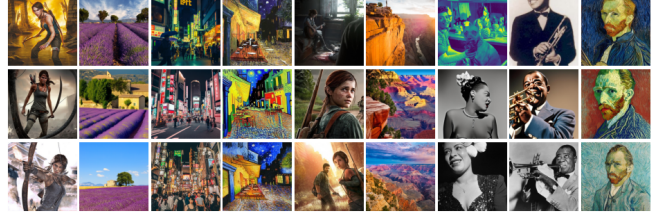
Image sampling strategy: We randomly sampled 5,000 image-caption pairs from the LAION-12M dataset and extracted their captions. Using Stable Diffusion 2, we then generated 5,000 images by employing the captions of the sampled pairs as prompts for each generation. The model was run using 30 inference steps and a guidance scale of 7.5.

For efficiency, we focus on only those captions that result in images with a greater than 0.85 similarity score when compared to the training images— Figure 2 presents a few of such images. This results in a set of 67 captions. For each of these 67 captions, we then apply different prompt engineering strategy to create new prompts. For each prompt, we generated 75 images with different random initializations to incorporate the inherent randomness in image generations [11]. The choice of 75 generations was a compromise between [4], who used 500 generations per image, and [11], who empirically found that 20–30 random initializations already provide a reasonably good representation of the diversity in generated images.

As we repeated the same process for all 67 captions, this approach resulted in 5,025 (67×75) generations per prompt engineering strategy. Creating samples for all three different prompt engineering strategies and one baseline set without prompt engineering led to a total of 20,100 generated images. We consider this number to be a healthy compromise between computational costs and representativeness. The structure ensured a comprehensive exploration of the effects of prompt engineering strategies while accounting for the inherent randomness within visual diffusion's image generation process.
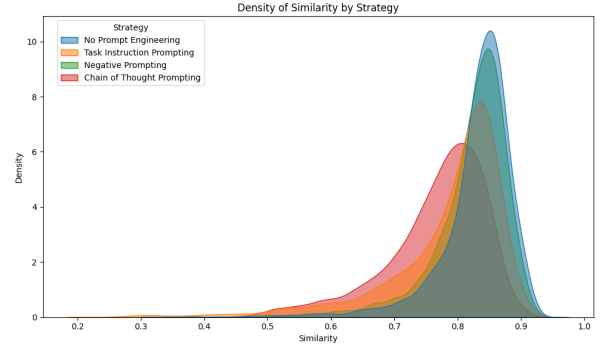
3.2 Similarity to training data

Comparison of distributions: Figure 3 displays the distribution of the similarity scores for each prompt engineering strategy using all generated images. Compared to task instruction prompting and chain-of-thought prompting, the distribution for the baseline approach peaks for higher similarity scores (around the threshold of

Figure 2: Examples of generated images exceeding the similarity threshold of 0.85. The top row displays the original images, whose captions were used as prompts for generating the images in the second row. The bottom row shows the most similar image from the LAION 12M dataset, identified based on cosine similarity of CLIP embeddings.

0.85 used in initial selection). This indicates that without prompt modifications, generated images tend to exhibit higher similarity to training data. In contrast, negative prompting had little effect on the distribution, with its curve closely resembling the baseline. Our results for negative prompting align with the observations made in [9] where negative prompting had little effect on the output images.

Figure 3: Density of the similarity score of all generations per applied prompt engineering strategy.

Statistical significance: To assess whether different prompt engineering strategies significantly influenced the similarity scores of generated images, we performed pairwise Wilcoxon rank sum tests on both average similarity and maximum similarity scores across prompt IDs. The Wilcoxon test was chosen because it is nonparametric and unlike the Student-t test, does not assume normality [29]. This is desirable because the distribution of similarity scores usually does not follow a normal distribution. Each test was performed at the prompt ID level to account for dependencies within the prompts while evaluating systematic differences between strategies. The results indicate that the differences between the task instruction prompting and the chain-of-thought prompt each compared to the baseline were statistically significant for both average and maximum similarity scores. However, the effect of negative prompting was not significant. The results, including the t statistics and the p values, can be found in Table 2.

3.3 Reduction in IP-infringement

As discussed in Section 2.2, a higher similarity score than 0.85 could be considered as high risk of IP infringement. Hence, we focus on the likelihood of generating the images exceeding this similarity threshold. We first consider the results over the entire set of 20, 100 images.

Table 2: Pairwise Wilcoxon rank-sum test results for average and maximum similarity scores across prompt strategies. Significant differences are indicated as follows: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Evaluated Strategies	Average Similarity		Maximum Similarity	
	t-statistic	p-value	t-statistic	p-value
No Prompt Eng. vs Task Instr. Prompting	4.3190	0.0000***	4.1649	0.0001***
No Prompt Eng. vs Negative Prompting	1.1439	0.2547	0.9905	0.3238
No Prompt Eng. vs Chain of Thought Prompting	6.8986	0.0000***	5.0428	0.0000***

Table 3 highlights that prompt engineering is particularly effective in mitigating IP infringement risks. It reduces the fraction of generated images that are similar to the training data. Without any prompt engineering, a total of 2,082 images (41.4% of 20,100 generated images) exceeded the 0.85 similarity score. The most effective prompt engineering strategy, chain-of-thought prompting, reduced this number to 9.6%—a 76.7% reduction compared to the baseline—leaving a total of only 484 images above the threshold.

The results become even more striking when examining the mean similarity score per prompt, calculated across the 75 generations per prompt ID. Without prompt engineering, 21 mean similarity scores exceeded the 0.85 threshold. Negative prompting reduced this number to 16, while task instruction prompting lowered it further to 7. The most effective approach, chain-of-thought prompting, reduced this number by 95% to just one prompt within the test sample. This means that on average, only one prompt among the 67 tested prompts generates an image remained in the critical similarity range above 0.85.

Table 3: Frequency of generations that are highly similar to training data

(a) Generations with similarity score > 85%

Prompt Engineering Strategy	Count	Percentage
No Prompt Engineering	2082	41.43
Task Instruction Fine Tuning	1026	20.42
Negative Prompt Engineering	1751	34.85
Chain of Thought Prompting	484	9.63

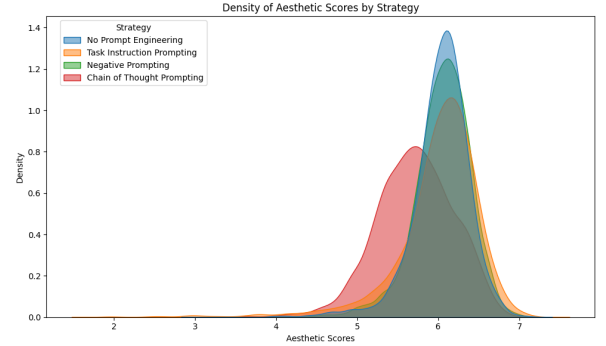**(b) Prompts for generations with average similarity score > 85%**

Prompting Strategy	Count	Percentage
No Prompt Engineering	21	31.34
Task Instruction Fine Tuning	7	10.45
Negative Prompt Engineering	16	23.88
Chain of Thought Prompting	1	1.49

3.4 Image Quality

To evaluate whether prompt engineering strategies for reducing copyright-related risks also impact the quality of generated images, we study two attributes: a) how asesthetic is the generated image ; and b) how well the generated images align with the original prompt.

Asthetic quality: For assessing the aesthetic quality of the generated images, we used the LAION-Aesthetics V2 predictor [22] to predict aesthetic scores. The estimated density plot of these aesthetic scores for each prompt engineering strategy is shown in Figure 4.

The plot indicates that images generated without prompt engineering and those using negative prompting tend to have the highest aesthetic scores, with their density peaks around similar values, 6.2 - 6.3; as a reference, images with an aesthetic score higher than 6 are considered to be reasonable [23, 20]. Task instruction prompting shows a slightly broader distribution, suggesting greater variability in aesthetic quality. In contrast, chain-of-thought prompting yields noticeably lower aesthetic scores, with a distribution shifted to the left and a broader spread compared to the other strategies. This indicates that while chain-of-thought prompting may be effective in mitigating copyright risks, it comes at the cost of generating images with lower aesthetic appeal.

Figure 4: Density of aesthetic scores of the generated images per prompt engineering strategy

Relevance to input prompts: To evaluate how well the generated images align with the requests, the prompts, we measured their semantic similarity to the original captions i.e., to the original captions that were extracted from the training dataset. This approach kept the underlying request for the assessment independent of the applied prompt engineering strategy. We assume that regardless of the applied prompt engineering strategy, the end-users expect the model output to meet the core of their request.
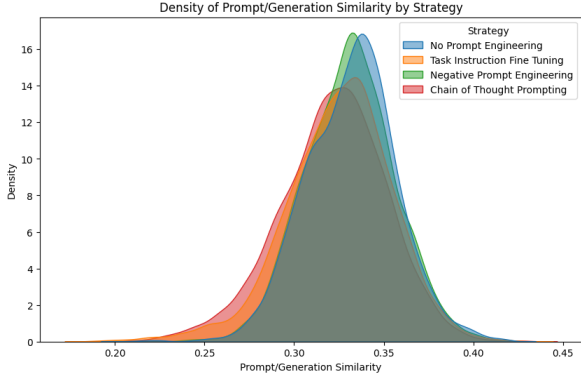
To conduct this evaluation, we calculated the CLIP similarity between the generated images and their original captions. To obtain CLIP embeddings for the captions, we applied the same CLIP model used for embedding the generated images. The similarity between the embeddings of the generated images and their corresponding captions was then calculated via cosine similarity.

Figure 5 presents the estimated density distribution of the CLIP similarity scores between the generated images and their corresponding prompts across different prompt engineering strategies. The results indicate that all strategies produce a similar overall distribution, with peaks centered around comparable similarity values. However, chain-of-thought prompting shows a slightly wider spread, which suggests more variability in how closely the generated images align with their original prompts. Negative prompting and no prompt engineering show the highest peak densities which implies more consistent alignment with the original captions. These findings suggest that while prompt engineering strategies influence generation outcomes, their impact on prompt adherence remains relatively limited.

3.5 Correlation between IP infringement risk reduction and image quality

To further explore how similarity to training data relates to image quality, we compute Pearson correlations [15] between the maximum CLIP similarity scores per prompt and two attributes: aesthetic

Figure 5: Density of CLIP similarity between the prompt and generated output per prompt engineering strategy.

score and prompt-caption similarity. The results are presented in Table 4. While chain-of-thought prompting yielded the lowest overall aesthetic score distribution, it exhibits the strongest positive correlation between similarity and both aesthetic quality ($r = 0.49$) and prompt adherence ($r = 0.33$). This indicates that, within this strategy, images with higher IP infringement risk tend to be more aesthetically pleasing and closely aligned with the original prompt. Task instruction prompting shows a moderate correlation with prompt adherence ($r = 0.25$), but only a weak correlation with aesthetic quality ($r = 0.13$), suggesting that higher-risk generations under this strategy tend to be better aligned with the prompt, though not necessarily more visually appealing. Other strategies show weaker correlations, indicating a less systematic relationship between similarity and image quality. Overall, these results highlight a trade-off: especially for chain-of-thought prompting, reducing IP risk may come at the cost of generating images that are less relevant and less aesthetically pleasing.

Table 4: Correlation between maximum similarity scores and performance evaluation parameters per strategy

(b) Correlation between Similarity Scores and Aesthetic Scores per Strategy

Prompt Engineering Strategy	Correlation
No Prompt Engineering	0.1352
Task Instruction Fine Tuning	0.1330
Negative Prompt Engineering	0.0772
Chain of Thought Prompting	0.4875

(a) Correlation between Similarity Scores and Prompt Adherence per Strategy

Prompting Engineering Strategy	Correlation
No Prompt Engineering	0.1462
Task Instruction Fine Tuning	0.2542
Negative Prompt Engineering	0.0850
Chain of Thought Prompting	0.3301

3.6   Practical recommendations

Since no particular prompting strategy outperforms all the others for all metrics, the choice of the prompting strategy is dictated by the choice of the metric, which in turn depends upon the application. We envision three different categories of applications and propose the ideal prompting strategy for each. The three application categories are: a) high risk; b) medium risk; and c) low risk. The higher the probability of being charged with an IP infringement lawsuit, the higher is the risk for an application. For example, a marketing campaign where the generated content is shared with a broad audience is high risk. In contrast, generated images used in a seminar held internally would be a low-risk scenario.

We have the following recommendation for different categories of applications:

- *High-risk applications:* we recommend the Chain-of-Through prompting strategy, which offers the greatest reduction in the IP infringement risk;
- *Medium-risk applications:* for medium-risk applications, Task Instruction prompting could be preferable. It strikes a good balance between the IP infringement risk and aesthetic/relevance score of the generated image.
- *Low-risk applications:* for low-risk applications, Task Instruction prompting and Negative prompting are recommended. These strategies provide the most relevant and aesthetically pleasing outputs while offering adequate safeguards against IP infringement.

4   Conclusions

We evaluate the effectiveness of prompting strategies in reducing the risk of IP infringement of visual GenAI by 3 metrics: similarity scores to training images, relevance and aesthetic value. Overall, we find that prompt engineering can reduce copyright-related risks in visual GenAI models, but its effectiveness varies depending on the chosen technique. Chain-of-thought prompting proved to be the most effective in IP risk mitigation. Negative prompting was the least effective strategy, while task instruction prompting yielded promising results in preventing training data replication while balancing this achievement with high aesthetic scores and strong prompt alignment. These strategies are particularly important given the legal uncertainty surrounding AI-generated content. We hope this work contributes to safer deployment practices and informs future standards for responsible visual generative AI use.

5   Limitations

Despite these promising results, several limitations remain. Prompt engineering alone cannot fully prevent copyright-infringing outputs, as certain high-risk prompts continue to yield recognizable replications. While our study provides a systematic framework for evaluating prompt-based mitigation strategies, future research should explore more sophisticated techniques, such as integrating model-side safeguards or hybrid approaches combining prompt engineering with interventions on all phases of a diffusion model's life cycle.

References

[1] dclure/laion-aesthetics-12m-umap · Datasets at Hugging Face.

[2] Sarah Andersen et al. Andersen v. stability ai, midjourney, deviantart, and runway ai, 2023. No. 3:23-cv-00201-WHO, U.S. District Court for the Northern District of California.

[3] Liam Barkley and Brink van der Merwe, 'Investigating the role of prompting and external tools in hallucination rates of large language models', *arXiv preprint arXiv:2410.19385*, (2024).

[4] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models, January 2023. arXiv:2301.13188 [cs].

[5] Avia Efrat and Omer Levy, 'The turking test: Can language models understand instructions?', *arXiv preprint arXiv:2010.11982*, (October 2020).

[6] U.S. District Court for the Northern District of California. Andersen v. stability ai, midjourney, and deviantart, 2024.

[7] Alvin Heng and Harold Soh, 'Selective amnesia: A continual learning approach to forgetting in deep generative models', *Advances in Neural Information Processing Systems*, **36**, (2024).

[8] Seunghoo Hong, Juhun Lee, and Simon S Woo, 'All but one: Surgical concept erasing with model preservation in text-to-image diffusion models', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21143–21151, (2024).

[9] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get What You Want, Not What You Don't: Image Content Suppression for Text-to-Image Diffusion Models, February 2024. arXiv:2402.05375 [cs].

[10] Vivian Liu and Lydia B Chilton, 'Design guidelines for prompt engineering text-to-image generative models', in *Proceedings of the 2022 CHI conference on human factors in computing systems*, pp. 1–23, (2022).

[11] Vivian Liu and Lydia B Chilton, 'Design Guidelines for Prompt Engineering Text-to-Image Generative Models', in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–23, New York, NY, USA, (April 2022). Association for Computing Machinery.

[12] Ning Miao, Yee Whye Teh, and Tom Rainforth, 'SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning', *arXiv preprint arXiv:2308.00436*, (October 2023).

[13] Microsoft. Customer copyright commitment required mitigations, 2024. Accessed: 2025-02-06.

[14] Michael D. Murray, 'Generative AI Art: Copyright Infringement and Fair Use', *SMU Science and Technology Law Review*, **26**(2), 259, (2023).

[15] Karl Pearson, 'Note on regression and inheritance in the case of two parents', *Proceedings of the Royal Society of London*, **58**, 240–242, (1895).

[16] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze, 'A self-supervised descriptor for image copy detection', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, (2022).

[17] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen, 'Reasoning with language model prompting: A survey', *arXiv preprint arXiv:2212.09597*, (December 2022).

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs].

[19] Cyrus Rashtchian, Charles Herrmann, Chun-Sung Ferng, Ayan Chakrabarti, Dilip Krishnan, Deqing Sun, Da-Cheng Juan, and Andrew Tomkins, 'Substance or style: What does your image embedding know?', *arXiv preprint arXiv:2307.05610*, (2023).

[20] Robin Rombach. stabilityai/stable-diffusion-2 · Hugging Face, 2022.

[21] Zahraa Al Sahili, Ioannis Patras, and Matthew Purver, 'FairCoT: Enhancing Fairness in Diffusion Models via Chain of Thought Reasoning of Multimodal Language Models', *arXiv preprint arXiv:2406.09070*, (October 2024).

[22] Christoph Schuhmann. Clip+mlp aesthetic score predictor, 2022.

[23] Christoph Schuhmann. LAION-Aesthetics | LAION, 2022.

[24] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, December 2022. arXiv:2212.03860 [cs].

[25] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and Mitigating Copying in Diffusion Models, May 2023. arXiv:2305.20086 [cs].

[26] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das, 'A comprehensive survey of hallucination mitigation techniques in large language models', *arXiv preprint arXiv:2401.01313*, (2024).

[27] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, Explaining, and Mitigating Memorization in Diffusion Models, July 2024. arXiv:2407.21720 [cs].

[28] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT, February 2023. arXiv:2302.11382.

[29] Frank Wilcoxon, 'Individual comparisons by ranking methods', *Biometrics Bulletin*, **1**(6), 80–83, (1945).

[30] Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and Lingpeng Kong. Diffusion of Thoughts: Chain-of-Thought Reasoning in Diffusion Language Models, December 2024. arXiv:2402.07754 [cs].

[31] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi, 'Forget-me-not: Learning to forget in text-to-image diffusion models', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1755–1764, (2024).