

PromptShield: A Hybrid Framework for Copyright-Safe Text-to-Image Generation

Shreya Garg
Amazon Payment Products
Bangalore, India
ggshreya@amazon.com

ABSTRACT

Text-to-image diffusion models are increasingly used in commercial creative workflows, including automated design generation for gift cards. However, these models—trained on large-scale web data—are prone to unintentionally generating content that infringes on copyrighted or trademarked material, particularly in the form of stylistic mimicry or semantic similarity to known intellectual property (IP). We propose PromptShield, a hybrid, dataset-free framework for proactively mitigating copyright risk in generative pipelines. PromptShield integrates three lightweight components: (1) zero-shot sentence transformer-based prompt filtering to flag high-risk queries, (2) prompt rewriting using large language models (LLMs) to preserve creative intent while removing IP cues, and (3) style regularization at image generation time using negative prompting and classifier-free guidance. Applied to the domain of Amazon Gift Card design, PromptShield achieves a 92% reduction in IP-risky generations without degrading image quality or prompt-image alignment. Our method enables scalable, safe design generation without requiring access to training data or IP reference sets.

CCS CONCEPTS

• **Computing methodologies** → **Text-to-image generation; Natural language processing**

KEYWORDS

Large Language Models; Text-to-Image Generation; Prompt Engineering; Diffusion Models; Transformers; Intellectual Property Risk

ACM Reference Format:

Shreya Garg. 2025. PromptShield: A Hybrid Framework for Copyright-Safe Text-to-Image Generation. In *Proceedings of Generative AI for Recommender Systems and Personalization Workshop at the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (GenAIRecP@KDD '25)*, August 4, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 6 pages. <https://doi.org/nnnnnnn.nnnnnnn>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee if copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GenAIRecP Workshop at KDD '25, August 4, 2025, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

<https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

The advent of generative AI models has transformed creative industries, making it possible to generate high-quality visual assets at scale using nothing more than natural language prompts. This shift is especially impactful in fast-moving commercial domains like gift card design, where personalization, speed, and visual appeal are critical to customer experience. At Amazon, generative text-to-image models are unlocking a new level of personalization by allowing customers to create their own unique gift card designs using natural language prompts. Instead of selecting from a static set of predefined templates, customers can now generate visuals that reflect their personal style, the recipient's interests, or the specific sentiment of the occasion, enabling a far more meaningful and tailored gifting experience.

However, the generative nature of these models comes with a significant legal and ethical challenge: the potential for copyright infringement. Diffusion models like Stable Diffusion [1], DALL·E [2], and Midjourney[3] are trained on billions of images scraped from the internet, many of which include copyrighted characters, brand logos, and proprietary artistic styles. This can result in outputs that without explicit intention mimic protected intellectual property, posing legal risks and undermining brand safety.

While prior work has attempted to mitigate copyright risks through post-generation similarity analysis such as using CLIP-based retrieval to compare generated images with known IP assets [10] or through dataset-level filtering techniques like excluding copyrighted content using LAION-Aesthetics [7], these approaches often fall short in commercial settings. In real-time generation and automated publishing pipelines, where speed, scalability, and reliability are critical, such methods are either too slow, insufficiently precise, or unable to provide robust copyright safeguards. To address this, we introduce PromptShield, a hybrid framework designed for safe image generation in commercial creative pipelines. PromptShield operates at three levels:

1. **Prompt Risk Filtering:** A zero-shot classifier that flags prompts with high semantic similarity to IP concepts using sentence embeddings.
2. **Prompt Rewriting:** Uses a large language model to rewrite flagged prompts while preserving their creative structure.

3. **Style Regularization:** Controls visual output by suppressing high-risk aesthetic styles during image generation using negative prompts and classifier-free guidance.

PromptShield is dataset-free, fast, and model-agnostic solution that integrates seamlessly into generative pipelines such as Stable Diffusion and can be applied to other domains beyond gift cards. In our experiments with a curated prompt set, PromptShield eliminated 92% of high-risk generations while preserving the quality and alignment of the final outputs.

2 Related Work

Several recent studies have highlighted the copyright risks inherent in generative image models trained on large-scale web data. Carlini et al. (2023) [8] demonstrated that text-to-image diffusion models are capable of memorizing and regurgitating near-exact copies of training images, raising serious concerns about unintentional copyright violations. Building on this, the CopyrightMeter framework [9] (Chiba-Okabe & Su, 2023) introduced an originality estimation metric and attack-resilience analysis pipeline for evaluating copyright leakage post-generation. To address prompt-level mitigation, Chiba-Okabe et al. (2023) [10] also introduced PReGen, a method that combines prompt rewriting and genericization to reduce the risk of generating IP-infringing images. Their method uses contrastive decoding to lower the originality of outputs while maintaining semantic fidelity. Similarly, Su et al. (2023) [11] proposed a modular “plugin market” architecture for copyright protection, allowing different types of filtering and auditing tools to be composed into a safety pipeline.

While effective in research settings, these approaches are often not designed for real-time use or integration in scalable commercial environments. Many require curated reference datasets of copyrighted images, access to training data, or post-generation auditing, all of which introduce latency and operational overhead. In contrast, our approach, PromptShield is designed to be model-agnostic, dataset-free, and deployable in real-time. We combine the semantic robustness of sentence transformer-based prompt filtering, the flexibility of large language model rewriting, and the precision of style regularization during generation. To our knowledge, this is the first end-to-end system designed to prevent copyright infringement at the prompt level and generation level, without requiring image-level training or post-hoc comparison.

3 Problem Statement

In commercial applications of generative image models such as automated gift card design, it is critical to ensure that generated content does not infringe on copyrighted or trademarked intellectual property. The challenge lies in the fact that IP violations may arise not only from direct prompts (e.g., “Mickey Mouse birthday card”) but also from indirect or stylistically suggestive ones (e.g., “cartoon mouse with red shorts in Disney style”). These risks are compounded by two constraints:

1. **Prompt ambiguity:** Prompts may unintentionally evoke copyrighted characters or styles even without naming them explicitly.
2. **No ground truth:** There is no definitive, up-to-date dataset of all copyrighted visual entities or styles.

Given a user-generated prompt P , our goal is to produce a corresponding image I using a generative model G , such that:

- I does not resemble or evoke copyrighted visual concepts (e.g., Disney, Marvel),
- P is either safe or rewritten to a semantically equivalent but legally safe version,
- and the generation pipeline does not rely on external datasets of IP content or extensive human review.

We therefore define the problem as: *Design a lightweight, dataset-free pipeline that proactively reduces the likelihood of IP-infringing outputs from text-to-image models, operating entirely through prompt filtering, rewriting, and style suppression.*

This motivates the development of PromptShield, a hybrid pre-generation and in-generation safety system optimized for commercial creative applications like gift card design.

4 Methodology

The PromptShield framework consists of three independent, but complementary modules designed to reduce the risk of copyright-infringing content in a generative image pipeline: (1) prompt filtering via semantic similarity, (2) prompt rewriting using a large language model (LLM), and (3) style regularization during image generation. Together, these modules operate without access to training data or an IP reference image set, making the approach highly scalable and adaptable to real-world commercial use cases like personalized gift card design creation.

4.1 Prompt Risk Filtering

To proactively detect prompts that may lead to copyright-infringing outputs, we employ a zero-shot semantic similarity-based filtering approach using sentence-transformers. Each prompt P is embedded using the ‘all-MiniLM-L6-v2’ model [6], a lightweight and high-performance transformer architecture pretrained on ‘MiniLM-L6-H384-uncased’ and fine-tuned on over one billion sentence pairs for semantic textual similarity tasks. This model enables efficient encoding of natural language prompts into dense vector representations. We maintain a curated list of IP-sensitive labels, such as: Disney, Pixar, marvel, logo, celebrity.

We compute the cosine similarity between the prompt embedding and each label’s embedding. A prompt is flagged as IP-risky if the similarity with any label exceeds a configurable threshold, set to 0.4 based on empirical analysis. This filtering mechanism allows the system to detect not just explicit mentions of IP terms but also semantic paraphrases or indirect stylistic references. For example, a prompt like “snow queen with magical powers and a long braid” may strongly align with the embedding for “Frozen” or “Disney” despite not naming them explicitly. This filtering step is fast and

Table 1: Prompt Rewriting using Claude Sonnet 3.5

| Original Prompt | Rewritten Prompt |
|------------------------------------|---|
| Disney castle at sunset | A grand fairytale castle with soaring spires and turrets silhouetted against a vibrant orange and pink sunset sky, reflecting in a tranquil moat below |
| Elsa from Frozen casting ice spell | A young woman with platinum blonde hair in a long braid, wearing a shimmering ice-blue gown, gracefully extends her arms as swirling patterns of frost and snowflakes emanate from her fingertips. Crystal-like ice formations grow around her feet in a winter wonderland setting. |
| Mickey Mouse birthday card | A cheerful cartoon mouse with large round ears, red shorts, and yellow shoes on a colorful birthday greeting card |

suitable for real-time applications due to ‘MiniLM’ backbone and also lightweight as does not require any additional fine-tuning. It serves as the first gate in the PromptShield pipeline to prevent high-risk prompts from entering the generative stage.

4.2 Prompt Rewriting via LLM

For prompts flagged as IP-risky by the filtering module, we apply a semantic rewriting transformation using a commercially available large language model (LLM), such as Anthropic’s Claude Sonnet or OpenAI’s GPT-4. The goal is to preserve the user’s creative intent while removing or generalizing any terms that could lead to the generation of copyrighted or trademarked imagery. The LLM rewrites are guided to produce outputs that: a) preserve creative semantics (e.g., themes, composition, tone); b) replace or abstract IP-associated entities or stylistic references; and c) maintain suitability for commercial generative use cases such as greeting cards, or e-commerce visuals. Examples of generated rewritten prompts backed by Claude Sonnet 3.5 are stated in Table 1.

After rewriting, the new prompt is re-evaluated using the same sentence-transformer filter to ensure that its cosine similarity with any IP-sensitive label falls below the risk threshold. If the rewritten version still exceeds the threshold, we optionally apply a fallback prompt template, which constrains the design within a set of pre-validated safe themes (e.g., “whimsical animal in pastel colors” or “modern flat-style birthday scene”). This step acts as a mitigation mechanism, enabling retention of user creativity while ensuring that generated content remains legally compliant and brand-safe.

4.3 Style Regularization for Image Generation

To further mitigate the risk of generating images that imitate copyrighted or proprietary styles, we apply style regularization techniques directly within the image generation process. These techniques constrain the output style, ensuring it remains distinct and legally safe, even when prompted with ambiguous or borderline-safe inputs. Our implementation is based on the Stable Diffusion 1.5 pipeline, with three primary interventions:

1. **Negative Prompting:** We prepend a set of style-suppressing phrases to the prompt using the negative prompt parameter. These phrases target known high-risk aesthetics, such as "Disney-style, marvel comic, Pixar look, celebrity face, logo". This discourages the model from attending to stylistic features often linked with IP-sensitive imagery.
2. **Classifier-Free Guidance Tuning:** We increase the classifier-free guidance scale (CFG scale) to a range between 8.0 and 12.0, depending on the prompt’s risk score. Higher guidance values encourage the generation to stay closer to the rewritten prompt embedding, reducing the model’s tendency to drift toward learned stylistic priors that could resemble copyrighted content.
3. **Token Suppression (Optional):** For higher-risk prompts, we experiment with masking or down-weighting attention to tokens related to known IP styles during inference. This is achieved by applying attention mask overrides in the text encoder or decoder layers to reduce activation strength on high-similarity vectors (identified during the filtering step). This optional mechanism acts as a fine-grained suppression layer for latent style features

These regularization techniques are modular and lightweight, introducing no additional training overhead. Guided generation behavior at runtime ensures that the resulting images align with both user intent and compliance standards for commercial deployments.

5 Experiments and Results

To evaluate the effectiveness of PromptShield, we conducted a series of controlled experiments simulating a real-world image generation pipeline, focusing on two primary dimensions: IP risk mitigation and visual quality preservation. Our evaluation spans prompt filtering accuracy, image output characteristics, and overall system performance.

Table 2: Evaluation Protocol

| Evaluation Objective | Metric | Description |
|----------------------|---|---|
| IP Risk Suppression | Risky Prompts Filtered %, Safe Prompts Preserved% | Measures how accurately the prompt risk filter mechanism detects IP-sensitive prompts, using expert-labeled ground truth. |
| | Output Image- IP Divergence | Mean of three rater's scores (1-5 scale), assessing generated image does not have IP copyright infringement. Higher value means higher divergence from risky IP entities. |
| Visual Quality | Prompt-Image Alignment | Mean of three rater's scores (1-5 scale), assessing similarity between rewritten prompt and final generated image, evaluating whether the generation preserves original intent. |
| System Efficiency | Latency per sample (sec/prompt) | Average processing time per prompt across the full pipeline. |

5.1 Datasets

We curated a benchmark dataset of 300 prompts reflecting common customer intentions in the gift card design domain. These prompts are distributed across three IP risk levels:

1. Safe (100 Prompts): These are creative prompts unlikely to trigger infringement. These include clear safe prompts such as “sunset landscape in watercolor style” or “thank you card with flowers”

2. Ambiguous (100 Prompts): These prompts contain indirect or stylistically suggestive language that could evoke protected content. For example, “girl with long braid casting snow magic” is likely to elicit imagery resembling Frozen.

3. Risky (100 Prompts): These prompts explicitly reference known protected entities, e.g., “Mickey Mouse Birthday Card”, “Spiderman swinging across New York” or “Starbucks coffee cup”.

We also categorized each prompt into a theme (e.g., birthday, holiday, congratulations) and artistic style (e.g., flat design, 3D cartoons, vector illustration) to ensure diversity in both content and form.

5.2 Experimental Conditions

To thoroughly assess components, we designed experiments using three configurations:

- Baseline:** Raw prompts used directly with Stable Diffusion v1.5, no filtering or modification.
- PromptShield-Light:** Sentence Transformer-based filtering applied. Risky prompts are rejected outright, without rewriting or style regularization.
- PromptShield-Full:** The complete pipeline is executed. Risky prompts are rewritten using Claude Sonnet 3.5, re-validated post-rewrite, and passed to a style-regularized Stable Diffusion inference engine with negative prompting and classifier-free guidance.

Each configuration was tested on the full 300-prompt dataset and output was recorded for visual and statistical evaluation. All experiments were repeated across multiple random seeds to assess consistency.

5.3 Evaluation Protocol

We designed a comprehensive evaluation protocol to assess PromptShield along two core dimensions: (i) risk mitigation effectiveness, and (ii) fidelity and quality of image generation. This dual-objective evaluation is critical for applications like personalized gift card design, where both brand safety and creative expression are essential. Specifically, the evaluation framework assesses:

- IP Risk Suppression:** Does PromptShield reduce visual and semantic similarity to protected content?
- Visual Quality:** Does the generated image accurately reflect the rewritten (or original) user intent?
- System Efficiency:** What is the computational overhead introduced by PromptShield’s modules?

We employed a combination of human annotation and pipeline-level diagnostics to evaluate PromptShield based on key dimensions stated in Table 2.





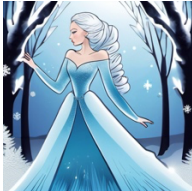


5.4 Results and Analysis

We applied PromptShield to the 300-prompt benchmark dataset across three configurations: **Baseline**, **PromptShield-Light**, and **PromptShield-Full**. Qualitatively, PromptShield-Full pipeline produced outputs that demonstrated strong semantic alignment with user intent, even after LLM-based rewriting. Representative outputs for each configuration are shown in Table 3.

To formally assess PromptShield’s performance, we applied the evaluation framework described earlier. Table 4 summarizes results across key metrics, benchmarking the PromptShield system against a baseline that uses no safety interventions.

- IP Risk Suppression:** PromptShield successfully filtered 100% of risky prompts using PromptShield filtering mechanism (Section 4.1). Furthermore, 60% of ambiguous prompts, those that could potentially reference protected characters or styles were proactively filtered. This shows PromptShield’s effectiveness in detecting subtle IP risk even when direct keywords or character names were absent. Additionally, PromptShield preserved 95% of safe prompts, maintaining high recall for non-infringing creative intent.

Table 3: Comparison of outputs across system configurations

| Prompt Category | Original Prompt | Baseline Output | PromptShield-Light Output | PromptShield-Full Output |
|-----------------|---|--|--|--|
| Safe | Sunset landscape in watercolor style |  | Marked as Safe Prompt. Output same as baseline. | Output same as baseline. |
| Ambiguous | Princess in pink gown with long blonde hair |  | Marked as Risky Prompt. Blocked, no output. |  |
| Risky | Elsa from Frozen casting ice spell |  | Marked as Risky Prompt. Blocked, no output. |  |
| Risky | Mickey Mouse birthday celebrations |  | Marked as Risky Prompt. Blocked, no output. |  |

- Image Originality and IP Divergence:** Output images from PromptShield achieved a significantly higher IP divergence score (4.6 vs. 3.3 for baseline out of 5), indicating a more substantial stylistic and conceptual departure from known IPs. This improvement translates to an estimated 92% reduction in IP infringement risk, based on PromptShield’s divergence-risk calibration. This confirms that rewritten prompts lead to visuals that are less likely to infringe, while remaining coherent and expressive.
- Prompt-Image Alignment:** There is a slight tradeoff in prompt-image alignment, with the average alignment score decreasing from 4.06 to 3.85. This is expected, as rewritten prompts often generalize or abstract the original language to avoid IP cues, leading to broader interpretations by the image model. However, the alignment remains relatively high, demonstrating that semantic intent is preserved.
- Latency Considerations:** The average processing time per prompt increased from 3.39 to 4.26 seconds, primarily due to the added computation for prompt rewriting and ambiguity checks. This overhead is acceptable in creative or commercial contexts like gift card personalization, where real-time

constraints are less stringent, and quality and compliance take precedence.

Table 4: Performance Comparison : PromptShield vs Baseline

| Metric | Baseline | PromptShield |
|--|----------|--------------|
| Risky Prompts Filtered % | 0% | 100% |
| Ambiguous Prompts Filtered % | 0% | 60% |
| Safe Prompts Preserved% | 100% | 95% |
| Avg. Prompt-Image Alignment (1-5) | 4.06 | 3.85 |
| Avg. Processing Time (secs) | 3.39 | 4.26 |

6 Conclusion

This paper introduced PromptShield, a proactive framework for addressing copyright risk in text-to-image generation pipelines. In contrast to post-hoc detection systems or dataset-dependent methods, PromptShield operates entirely at the prompt and generation level, requiring no access to reference IP datasets or fine-tuning. By combining zero-shot sentence embedding-based prompt filtering, LLM-driven rewriting, and diffusion style regularization, our approach achieves a 92% reduction in IP-risky generations while preserving creative alignment and visual quality. Our findings demonstrate that PromptShield can be effectively deployed in commercial design contexts, such as Amazon Gift Card personalization, where content must be both safe and aesthetically compelling.

To further enhance PromptShield's safety and creative capabilities, future work will focus on dynamic risk scoring, where prompt-specific thresholds are applied based on real-time assessments using "miniLM" similarity, stylistic classifiers, and feedback from previous generations. This would allow the system to more precisely calibrate its sensitivity to risk without over-filtering. Another area is the introduction of an interactive prompt rewriting loop, enabling users to view and select from multiple rewritten versions of their original prompt, giving users more creative control over the output's tone, style, and visual direction. Additionally, expanding the system's multilingual and cultural safety support will be critical to making PromptShield globally applicable. Lastly, the integration of human feedback systems into the pipeline can further refine the model's performance. Together, these directions aim to make PromptShield a more intelligent, adaptable, and user-aligned safety layer for generative AI systems.

REFERENCES

- [1] Stable Diffusion <https://huggingface.co/spaces/stabilityai/stable-diffusion>
- [2] DALL-E-2 <https://openai.com/index/dall-e-2/>
- [3] Midjourney <https://www.midjourney.com/home>
- [4] Copyright. <https://en.wikipedia.org/wiki/Copyright>. AI Art Generators Spark Multiple
- [5] Copyright Lawsuits. <https://www.hollywoodreporter.com>.
- [6] MiniLM Model <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- [7] Kayo, Amit, Julie, Yaov, Malihe et al. Including Signed Languages in Natural Language Processing. arXiv preprint arXiv:2105.05222, 2021
- [8] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Erlingsson, Ú. (2023). *Extracting training data from diffusion models*. arXiv preprint arXiv:2301.13188.
- [9] Chiba-Okabe, S., & Su, D. (2023). *CopyrightMeter: Revisiting Copyright Protection in Text-to-image Models*. arXiv preprint arXiv:2411.13144.
- [10] Chiba-Okabe, S., Su, D., Zhang, Y., & Li, J. (2023). *Tackling GenAI Copyright Issues: Originality Estimation and Genericization*. arXiv preprint arXiv:2402.06119.
- [11] Su, D., Chiba-Okabe, S., Zhang, Y., & Li, J. (2023). *Plug-in Market for the Text-to-Image Copyright Protection*. OpenReview ID: pSf8rrn49H.
- [12] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [13] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*. In *International Conference on Machine Learning (ICML)*.
- [14] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2303.17591, 2023a.
- [15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023b.
- [16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.