Guardians of Generation: Dynamic Inference-Time Copyright Shielding with Adaptive Guidance for AI Image Generation

March 10, 2025

Abstract

Modern text-to-image generative models can inadvertently reproduce copyrighted content memorized in their training data, raising serious concerns about potential copyright infringement. We introduce Guardians of Generation, a model-agnostic inference-time framework for dynamic copyright shielding in AI image generation. Our approach requires no retraining or modification of the generative model's weights, instead integrating seamlessly with existing diffusion pipelines. It augments the generation process with an adaptive guidance mechanism comprising three components: a detection module, a prompt rewriting module, and a guidance adjustment module. The detection module monitors user prompts and intermediate generation steps to identify features indicative of copyrighted content before they manifest in the final output. If such content is detected, the prompt rewriting mechanism dynamically transforms the user's prompt—sanitizing or replacing references that could trigger copyrighted material while preserving the prompt's intended semantics. The adaptive guidance module adaptively steers the diffusion process away from flagged content by modulating the model's sampling trajectory. Together, these components form a robust shield that enables a tunable balance between preserving creative fidelity and ensuring copyright compliance. We validate our method on a variety of generative models (Stable Diffusion, SDXL, Flux), demonstrating substantial reductions in copyrighted content generation with negligible impact on output fidelity or alignment with user intent. This work provides a practical, plug-and-play safeguard for generative image models, enabling more responsible deployment under real-world copyright constraints.

1. Introduction

Advances in text-to-image diffusion models have revolutionized image generation, enabling users to transform descriptive text into high-fidelity visual outputs [18, 30, 32] and opening new avenues for creative expression. However, alongside these impressive capabilities come significant ethical and legal challenges. Recent studies have shown that even subtle or indirect prompts can inadvertently generate images that closely resemble copyrighted or trademarked works [16, 40], raising serious concerns about compliance, intellectual property rights, and potential legal liabilities [1–4, 6].

Existing safeguards adopt a variety of strategies to mitigate copyright infringement. Some rely on watermarking, which seeks to embed distinctive signatures in either training data or generated outputs [12, 42]. Although watermarks can visibly flag ownership, they do not necessarily avert infringing imagery in the first place, and adversaries may attempt to remove or obscure these marks. Dataset-level filtering attempts to preemptively remove protected content during model training [24], but this approach is impractical to apply repeatedly and often training data may not be available. Other studies have explored reinforcement learning to steer diffusion processes away from protected imagery [36], but these methods require extensive retraining, limiting their large-scale applicability. Concept erasure methods [14, 35] seek to remove protected concepts at the model level, yet they are computationally expensive and ill-suited for addressing fine-grained copyright protection requirements. Further, these solutions often struggle to address ad-hoc prompts that deliberately or indirectly reference copyrighted entities or stylistic traits.

At the user interface level, prompt rewriting has emerged as a complementary approach to sanitize inputs before they reach the generation pipeline [16]. While effective at removing explicit references, these methods can be bypassed by malicious or creatively phrased prompts that evade simple keyword checks. Furthermore, aggressive rewriting may distort the user's original intent, undermining the expressive power of text-to-image models. These challenges underscore a fundamental tension: preserving the rich creativity of generated content while robustly preventing the synthesis of infringing material.

In this paper, we propose Guardians of Generation

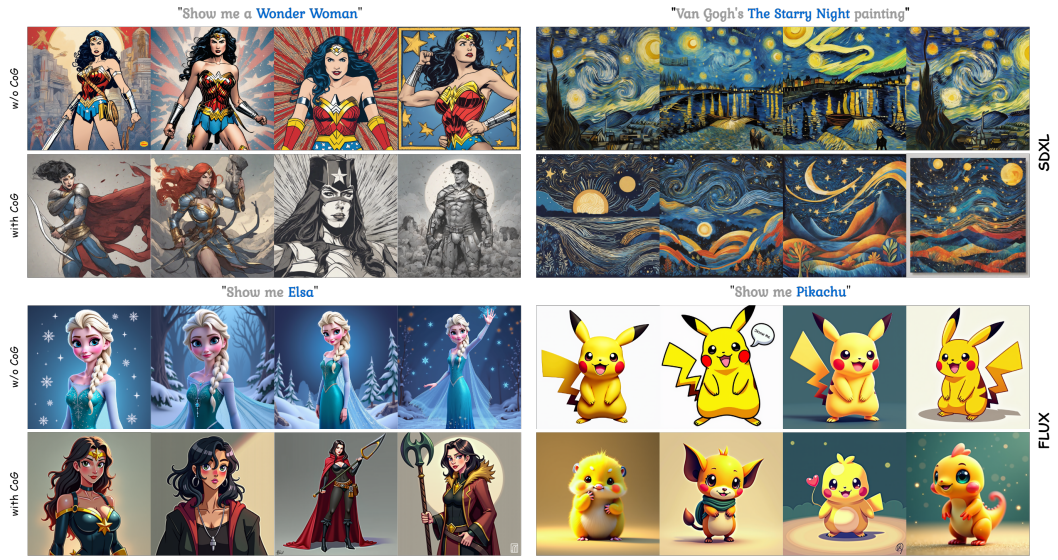

Figure 1. Image generation results for Flux and SDXL models, with and without our **GoG** copyright protection. Each row corresponds to a distinct input prompt, demonstrating that our pipeline preserves image quality and creative intent while ensuring strict copyright compliance.

Table 1. A comparative analysis of key attributes across various copyright protection strategies in text-to-image diffusion models, showing the advantages of the proposed GoG framework over unlearning, watermarking, and style cloaking methods

Attribute	Unlearning based	Watermark based	Style cloaking	**GoG (ours)**
No model retraining?	✗	✓	✓	✓
Inference-time support?	✗	✓	✗	✓
Support for newly emergent copyright concerns?	✗	✗	✗	✓
Preserves original model parameters?	✗	✓	✓	✓
User control over mixing / style?	✗	✗	✗	✓
Fidelity of the model retained?	✗	✗	✓	✓

(GoG), a unified pipeline for copyright protection at inference time—without requiring any model retraining. Our approach first detects protected or trademarked references in the user prompt using an embedding-based similarity check paired with a language-model disambiguation step [13]. Flagged prompts are then rewritten by a large language model (LLM)[25, 37] to remove explicit or subtle references to the targeted content. Next, both the original and sanitized prompts are integrated in a single diffusion pass via an adaptive Classifier-Free Guidance (CFG) mechanism, enabling a tunable balance between preserving core user intent and diluting infringing elements (see Figure1). Unlike watermarking or dataset-level filtering, our method directly prevents infringing content from emerging without altering model weights (see Table 1). Our model-agnostic pipeline achieves consistent performance across multiple architectures (Stable Diffusion 2.1, Stable Diffusion XL, Flux), as extensive experiments demonstrate robust mitigation against trademarked content while retaining the stylistic and thematic essence of the original prompt. The contributions of this paper are as follows:

Prompt Detection and Rewriting. We develop an embedding-based detection framework that flags suspicious concepts, augmented by an LLM disambiguation routine. Detected prompts are then automatically sanitized to eliminate direct or subtle references.

Adaptive CFG. We introduce an extension of Classifier-Free Guidance that combines the user's original and sanitized prompts, granting a tunable "mixing weight" to balance creative fidelity and legal compliance.

Model-Agnostic Implementation. Our pipeline seamlessly supports different diffusion platforms (SD 2.1, SDXL, Flux) without retraining or modifying their internal weights, reflecting the practical feasibility of adopting our method in real-world deployments.

Extensive Evaluation. We benchmark our system on a diverse set of concepts and prompts, including indirect anchoring and complex prompts to replicate copyrighted images. We demonstrate superior prevention rates compared to the existing state-of-the-art methods.

2. Background

Copyright Liability in AI Image Generation: *An Enterprise View.* In text-to-image generative models, copyright

liability depends on whether the generated output reproduces protected elements of a source work. Direct copying—where an AI output is nearly identical to the original, as seen in Thomson Reuters v. ROSS Intelligence Inc.[9], can lead to infringement claims, exposing enterprises to significant legal and financial risks. In contrast, if an AI-generated image only evokes the overall style or thematic elements, it may be deemed a derivative work, a classification discussed in legal and academic analyses[8, 22]. The intended use of the output—whether for internal analytics, customer-facing applications, or commercial marketing, further influences liability, with courts more likely to find fair use in cases of transformative, generic, or de minimis reproductions [5, 6]. Comparative jurisdictional rulings add further nuance: for example, the Beijing Internet Court has granted copyright protection when substantial human input is evident [1], whereas the Hangzhou Internet Court and a 2024 Czech ruling impose liability when outputs closely mirror protected training data or lack clear human authorship [2, 3]. Supplementary analyses by the RAND Corporation and comprehensive overviews available on Wikipedia further highlight the multifaceted legal landscape enterprises must navigate [4, 7].

Related work. The powerful capabilities of generative models pose pressing concerns around the unauthorized reproduction of copyrighted or trademarked materials [11, 21]. Even ostensibly innocuous or indirect prompts can yield outputs that closely mimic protected works, raising ethical and legal dilemmas [17, 31]. He *et al.* [16] show that certain copyrighted characters can be reproduced by text-to-image diffusion models, sometimes triggered by minimal keyword prompts. In contrast to prior research focusing on either broad memorization issues [11, 39] or single-model interventions, they develop an evaluation framework that systematically tests a range of generative models, highlighting both the subtlety and pervasiveness of copyright infringements. Their experiments emphasize that naive guardrails, such as prompt rewriting or negative prompts, frequently fail to eliminate all traces of infringing elements. A common strategy is *watermarking training images* [43] or generated outputs [12] to signal ownership. While effective for labeling, these methods don't prevent generation and can be circumvented. Similarly, large-scale dataset curation (e.g., removing copyrighted samples [33]) is limited by content volume and the unpredictability of future infringements.

Some studies propose *rewriting or filtering prompts before generation*. He *et al.* [16] identify "indirect anchors" and cleanse them via negative or altered prompts, while decoding-time strategies [15] adjust sampling to avoid protected content. However, these methods often rely on simple keyword matching and can be defeated by subtle rephrasings [19]. In contrast, our approach combines an embedding-based detector with an LLM-based rewrite, integrating both sanitized and original prompts via adaptive guidance.

Model-level approaches (unlearning, model editing) also address copyright risks. Ko *et al.* [20] show that unlearning specific concepts degrades alignment and propose boosting methods to preserve quality, while Qiu *et al.* [27] introduce orthogonal finetuning to adapt model weights without losing semantics. Other works explore image-to-image unlearning [38] and local conditional controlling [41] to modify specific regions. However, these techniques typically require retraining or fine-tuning, reducing their flexibility for on-demand enforcement. Diffusion models can *mimic the styles of living artists* [34], raising copyright and moral rights concerns. Such tools disrupt style imitation during training with subtle perturbations, but these methods modify data or the training process rather than providing inference-time defenses.

Collectively, these works highlight the need for robust, user-facing controls that avoid costly re-training. [16] shows that copyright vulnerabilities persist across major text-to-image systems. Our *model-agnostic* pipeline operates purely at inference by combining an embedding-based protected concept detector, LLM rewriting, and adaptive Classifier-Free Guidance, addressing the limitations of watermarking and naive rewriting without the overhead of unlearning or fine-tuning.

3. Guardians of Generation

We propose Guardians of Generation (GoG) as a three-stage pipeline: *protected concept detection*, *prompt rewriting*, and *adaptive classifier-free guidance*, to transform a potentially policy-violating prompt into a safe yet semantically faithful text-to-image generation (see Figure 2). The detailed step-by-step procedure is outlined in Algorithm 1. Below we detail each stage, highlighting how they jointly ensure copyright protection and semantic preservation.

3.1. Protected Concept Detection

Let p denote the user prompt and $C = \{c_1, \ldots, c_m\}$ be the set of protected concepts (pre-defined by policy). We employ two complementary detectors: an *embedding-based similarity filter* and an *LLM-based policy judge*. We compute the embedding $f_{\text{emb}}(p)$ using a pre-trained encoder and similarly obtain $f_{\text{emb}}(c_i)$ for each concept (with its possible synonyms: semantic and entity relationship). The cosine similarity is given by:

$$s_i = \frac{\langle f_{\text{emb}}(p), f_{\text{emb}}(c_i) \rangle}{\|f_{\text{emb}}(p)\| \, \|f_{\text{emb}}(c_i)\|}, \quad i = 1, \ldots, m. \tag{1}$$

If any $s_i \geq \tau$ (set based on policy strictness), the prompt is flagged.

Algorithm 1 Guardians of Generation (GoG)

Require:
p: user prompt; $\mathcal{C}$: set of protected concepts with synonyms; τ: similarity threshold; f_θ: diffusion model with CFG support; α: adaptive CFG mixing weight; η: guidance scale

Ensure:
I: generated image without protected elements

Concept Detection:
$\mathcal{F} \leftarrow \text{Detector}(p, \mathcal{C}, \tau)$ ▷ Identify protected concepts
$p_{re} \leftarrow \begin{cases} \text{LLMRewrite}(p, \mathcal{F}) & \text{if } \mathcal{F} \neq \emptyset \\ p & \text{otherwise} \end{cases}$
Embedding Calculation:
$(\phi_p, \phi_{p_{re}} \leftarrow \text{EncodePrompts}(p, p_{re})$
$\phi_{\text{mix}} \leftarrow (1-\alpha)\phi_p + \alpha\phi(p_{re})$ ▷ Linear interpolation
Diffusion Sampling:
$z_T \sim \mathcal{N}(0, I)$
for $t = T$ **down to** 1 **do**
  $\epsilon_{\text{uncond}} \leftarrow f_\theta(z_t, \phi_{neg}, t)$
  $\epsilon_{\text{cond}} \leftarrow f_\theta(z_t, \phi_{\text{mix}}, t)$
  $\widehat{\epsilon}_t \leftarrow \epsilon_{\text{uncond}} + \eta(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$
  $z_{t-1} \leftarrow \text{SchedulerStep}(z_t, \widehat{\epsilon}_t, t)$
end for
$I \leftarrow \text{VAEdecode}(z_0)$
return I

Not all problematic prompts contain keywords that are easy to catch with embeddings; some require contextual understanding. We leverage a f_{LLM} (e.g., GPT-4) to judge the prompt against the content policy. The LLM is prompted with the text of p along with policy descriptions Π, and it outputs a judgment score or label indicating if p violates any rule. The LLM's extensive knowledge allows it to interpret nuanced or implicit content and apply complex policy rules. For example, it can flag a prompt that subtly requests disallowed content even if specific banned terms are absent. We combine these signals to decide if p contains protected concepts:

$$I_{\text{flag}}(p) = \mathbb{1}\Big\{ \max_i s_i > \tau \ \vee \ f_{\text{LLM}}(p, \Pi) = 1 \Big\}. \quad (2)$$

If $I_{\text{flag}}(p) = 1$, the prompt is forwarded for rewriting.

3.2. Prompt Rewriting

Once a protected concept is detected in p, we invoke an LLM-based prompt rewriting to obtain a sanitized prompt p_{re}. The goal is to remove or replace the disallowed elements of p while preserving the prompt's high-level semantics and intent. We frame this as generating a new prompt that maximizes semantic similarity to the original concept under a copyright constraint. Let $D(p)$ denote the set of detected disallowed concepts in p. The rewriting process can

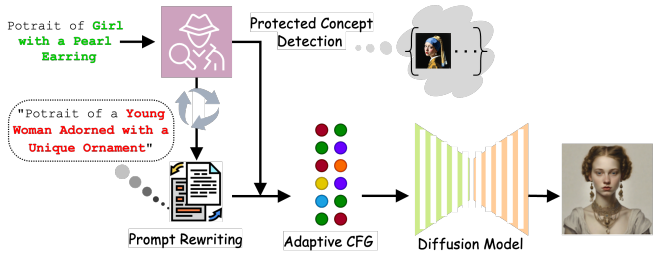

Figure 2. Overview of the complete Guardians of Generation (GoG) copyright protection pipeline.

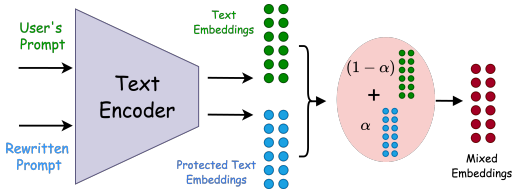

Figure 3. Adaptive CFG with mixture of embeddings

be conceptualized as solving a constrained optimization:

$$p_{\text{re}} = \arg\max_Q \ \text{Sim}(Q, p) \quad \text{s.t.} \quad D(p) \not\subset Q.$$

where $\text{Sim}(Q, p)$ is a semantic similarity measure between prompts and the constraint requires that Q contains none of the disallowed concepts identified in p. Formally, we denote:

$$p_{\text{re}} = \mathcal{R}\big(p,\ D(p)\big), \quad (3)$$

where $\mathcal{R}$ is an LLM-driven transformation. The sanitized prompt is iteratively verified until it contains no disallowed elements. For example, if p asked for "`portrait of a girl with pearl earrings`," and *pearl earrings* is disallowed, the LLM might output "`a portrait of a young woman adorned with a unique ornament`," preserving the intended person (a girl) but abstracting away the specific object. By leveraging the LLM's rich understanding of language, the rewriting step can intelligently fill gaps or alter descriptions so that the resulting prompt p_{re} remains coherent and faithful to the user's request, minus the protected elements.

It's worth noting that the LLM's rewrite favors maintaining all allowed aspects of the prompt (style, composition, attributes, etc.) and only modifies what is necessary to comply with policy. The high-level intent – what the user essentially wants to see – is thus made explicit in p_{re} without hidden copyrighted elements. This rewritten prompt will guide the image generation in the next stage.

Iterative Safeguards. In some cases, the LLM may inadvertently reintroduce new or tangential references that conflict with the protected set C. To mitigate this,

Figure 4. Effect of increasing the mixing weight (α) on image generation, ranging from 0 (no copyright protection) to 1 (fully protected generation). At $\alpha = 0$, the model follows the original user prompt without protection, potentially generating copyrighted content. As α increases, the generated images gradually deviate from the copyrighted concept while preserving some characteristics. Beyond $\alpha = 0.7$, **GoG** effectively avoids copyright violations while still adhering to the intent of the initial prompt.

the rewritten prompt p_{re} is re-evaluated by the detection pipeline, and the rewriting process is repeated until no elements from C remain. The final sanitized prompt p_{re} is then forwarded to the generation pipeline, where it is combined with the original prompt p using an adaptive classifier-free guidance mechanism.

3.3. Adaptive Classifier-Free Guidance

In the final stage of our framework, we seamlessly integrate both the original prompt p and its sanitized counterpart p_{re} into the image generation process (see Figure 3). This is accomplished through an adaptive classifier-free guidance mechanism that ensures the generated images remain faithful to the user's creative intent while strictly adhering to copyright constraints.

Mixture of Embeddings: For each text encoder E in the diffusion pipeline, whether it is the single encoder used in SD 2.1 or the dual encoder setup in SDXL and Flux, we compute the embeddings for both the original prompt and its rewritten version $\phi_p = E(p)$ and $\phi_{p_{\text{re}}} = E(p_{\text{re}})$. We then blend these embeddings using a linear interpolation governed by a tunable mixing weight $\alpha \in [0, 1]$:

$$\phi_{\text{mix}} = (1 - \alpha)\, \phi_p + \alpha\, \phi_{p_{\text{re}}} \tag{4}$$

This mixture allows us to modulate the influence of the sanitized prompt relative to the original, ensuring that while the protected content is suppressed, the overall semantic details of the prompt are preserved. The mixing weight α and the guidance scale η (of CFG) jointly modulate the balance between sanitized and original semantic features, ensuring that modifications do not dilute the stylistic and thematic nuances of the prompt.

Noise Prediction and Latent Update: Following the principles of classifier-free guidance [18], the diffusion model f_θ is executed twice at each denoising step. First, an unconditional noise estimate ϵ_{uncond} is computed by providing either an empty or negative embedding ϕ_{neg}. Next, the conditional noise estimate ϵ_{cond} is obtained using the mixed embedding ϕ_{mix}: $\epsilon_{\text{cond}} = f_\theta(x_t, \phi_{\text{mix}}, t)$ and $\epsilon_{\text{uncond}} = f_\theta(x_t, \phi_{\text{neg}}, t)$. The final noise prediction is then formed by interpolating between these two estimates:

$$\epsilon_t = \epsilon_{\text{uncond}} + \eta\Big(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}}\Big) \tag{5}$$

where η is the guidance scale that controls the overall strength of the conditioning. Together, these steps form an adaptive guidance strategy that elegantly blends the copyright-protected rewritten prompt with the rich semantic details of the original prompt. This ensures that the generated images are both compliant with copyright constraints and faithful to the user's creative vision.

Model-Agnostic Implementation: Our framework is designed to be agnostic to the underlying diffusion architecture. All three instantiations SD 2.1, SDXL, and Flux follow the same core procedure. While the computation of prompt embeddings and the structuring of latent representations vary, the fundamental classifier-free guidance (CFG) equation in Eq. (5) remains unchanged. In particular, SD 2.1 employs a single text encoder (CLIP [28]), whereas SDXL and Flux utilize dual text encoders (CLIP and T5 [29]) with the same input prompt p provided to each encoder. This modular design enables seamless integration of our method with various state-of-the-art diffusion models without requiring modifications to their internal archi-

η	SSIM	LPIPS	CLIP-I	CLIP-T
2.0	0.16	0.65	0.66	0.17
3.0	0.18	0.66	0.66	**0.17**
4.0	0.20	0.65	0.67	0.16
5.0	0.21	0.67	0.68	0.16
6.0	0.21	0.66	0.69	0.16
7.0	0.22	0.66	0.69	0.16
8.0	**0.23**	**0.67**	**0.70**	0.16

Table 2. **GoG** performance on SD 2.1 at different guidance scale η and mixing weight $\alpha = 0.5$

η	α	SSIM	LPIPS	CLIP-I	CLIP-T
2.0	0.5	0.21	0.55	0.78	0.17
	0.7	0.22	0.56	0.76	0.17
3.0	0.5	**0.22**	0.56	0.80	**0.17**
	0.7	0.22	0.56	0.76	0.17
4.0	0.5	0.21	0.55	0.81	0.17
	0.7	0.21	0.57	0.77	0.17
5.0	0.5	0.21	0.57	0.82	0.17
	0.7	0.21	0.57	0.77	0.16
6.0	0.5	0.2	0.57	0.82	0.16
	0.7	0.20	0.58	0.77	0.16
7.0	0.5	0.19	0.57	0.83	0.16
	0.7	0.19	0.58	0.77	0.16
8.0	0.5	0.19	0.57	**0.83**	0.16
	0.7	0.18	**0.59**	0.79	0.16

Table 3. **GoG** performance on SDXL at different guidance scale η and mixing weights α

tectures.

4. Experiments and Results

Experiment Settings. We perform experiments on three text-to-image generative models: Stable Diffusion 2.1 [30], Stable Diffusion XL [26], and Flux [10]. To evaluate copyright protection, we assemble a diverse set of 33 protected concepts spanning various categories, including movie characters, animated figures, video game protagonists, brand logos, portraits of actors and singers, art styles, and famous paintings. For each concept, we establish semantic and entity relationships by incorporating synonyms and related references. Additionally, three prompt variants of different lengths are generated per concept (see Table 7). For each prompt, we sampled 4 images across 7 different guidance scales.

Metrics We evaluate the degree of copyright protection with six metrics CLIP-I, CLIP-T, LPIPS, SSIM, CONS, DETECT [16]. When evaluating the effectiveness of copyright protection in AI-generated images, it is crucial to interpret similarity metrics within a balanced, intermediate range. Metrics such as CLIP-I, CLIP-T, and LPIPS quantify semantic and perceptual alignment between the generated image, the original copyrighted image, and the user's textual prompt. Extremely high values of these metrics suggest minimal modifications, risking copyright infringement, while excessively low values indicate significant deviation from the intended visual style or user request, diminishing utility. Thus, for effective copyright protection that also partially satisfies the user's intent, these metrics should ideally lie within a moderate, balanced range. We also employ CONS, a VQA model [23] that checks for key visual features in the generated image, and DETECT, which counts occurrences of target entities to measure unintended replication [16].

η	α	SSIM	LPIPS	CLIP-I	CLIP-T
2.0	0.5	0.36	0.55	0.85	0.15
	0.65	0.36	0.56	0.79	0.15
3.0	0.5	**0.36**	0.54	**0.85**	**0.16**
	0.65	0.35	0.57	0.80	0.15
4.0	0.5	0.34	0.55	0.84	0.16
	0.65	0.33	0.57	0.79	0.15
5.0	0.5	0.33	0.56	0.82	0.16
	0.65	0.32	**0.59**	0.78	0.15

Table 4. **GoG** performance on FLUX at different guidance scale η and mixing weights α

	SDXL			FLUX		
η	$CONS_{unpro}$	$CONS_{\alpha=0.7}$	DETECT	$CONS_{unpro}$	$CONS_{\alpha=0.65}$	DETECT
2.0	0.46	0.44	1	0.42	0.47	7
3.0	0.43	0.50	1	0.41	0.44	4
4.0	0.46	0.46	2	0.41	0.44	4
5.0	0.46	0.47	4	0.41	0.46	3
6.0	0.42	0.48	2	0.39	0.46	3
7.0	0.44	0.46	9	0.38	0.42	6
8.0	0.43	0.44	1	0.42	0.47	4

Table 5. CONS and DETECT scores on Flux and SDXL models. We compare the CONS metric with and without GoG protection.

Dataset We built a dataset across four domains: cartoon/animated characters (11 targets), famous movie stars/singers/people (8 targets), brand logos (8 targets), and famous paintings (6 targets). For each target, three prompts were generated, from simple (e.g., "Show me a Mario") to elaborate (e.g., "An intricate scene featuring Pikachu leading a group of diverse Pokémon through a challenging forest filled with obstacles and hidden Pokéballs"), to test how prompt complexity affects consistency and potential IP infringement (see Table 7 for prompt's length distribution in our dataset). Overall, the dataset includes 99 diverse prompts. For indirect anchoring, we focused on the 11 cartoon characters, using descriptive cues instead of explicit names. We observed that all three base models produced copyrighted images even when the target name was omitted. Following [16], although the outputs aligned with the prompts, the generated images were too similar to the original copyrighted concepts (see Figure 1 and Figure 6).

Balancing semantic fidelity with visual variation: Our evaluation indicates that the generated images maintain strong semantic fidelity with the user's prompt, as evidenced by consistent CLIP-T scores (0.15–0.17) across models. At the same time, perceptual and structural mod-

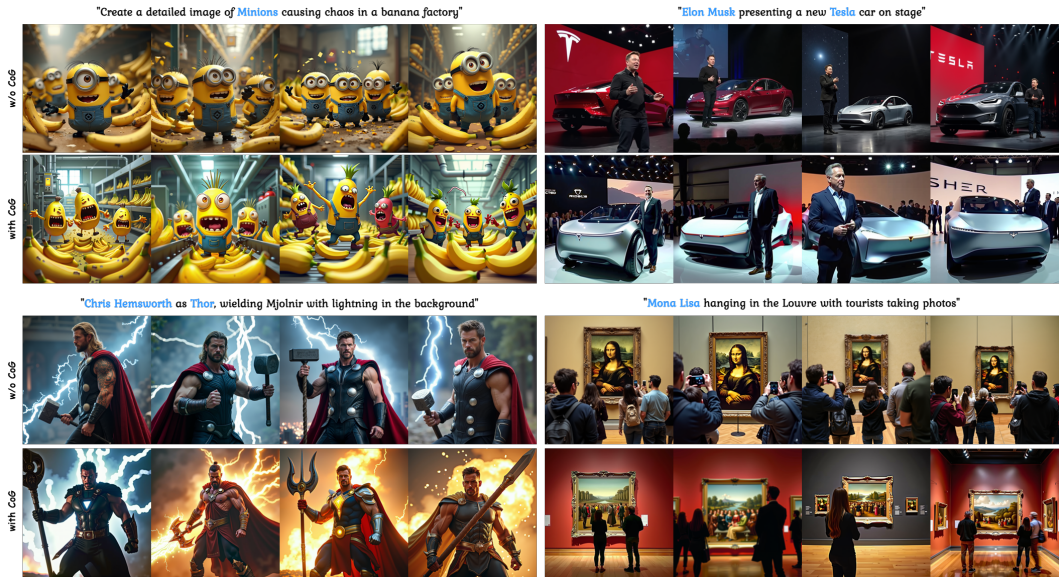

Figure 5. Effectiveness of **GoG** on complex prompts. Our approach protects the targeted concept while preserving the overall context and semantics of the original prompt, ensuring that unrelated elements remain unaffected.

Figure 6. We use indirect anchoring as a potential jailbreak mechanism. Without our **GoG**, the model produces copyrighted images even when the prompt lacks explicit references (1st row). In contrast, our approach prevents generating protected content while preserving the prompt's semantics (2nd row), underscoring its robustness against indirect anchoring vulnerabilities

ifications vary by model. For instance, LPIPS scores for SD2.1 and SDXL (0.16–0.22) (see Table 2 and Table 3) suggest high perceptual similarity, while Flux's higher LPIPS (0.32–0.36) (see Table 4) indicates a greater degree of visual deviation. SSIM results further show that SD2.1 (0.65–0.67) retains more of the original structure compared to SDXL and Flux (0.55–0.59), and CLIP-I values (ranging from 0.66–0.85) reflect a moderate image-to-image similarity that helps balance between retaining desired visual cues and avoiding excessive replication of copyrighted details.

Overall, these metrics suggest that GoG successfully achieves the desired balance: the outputs remain consistent with the intended semantic message while introducing sufficient perceptual and structural modifications to mitigate direct copying risks. In this context, although SD2.1 and SDXL produce images that are more visually similar to the originals, Flux's approach offers a greater degree of deviation. This outcome implies that the generated images align well with the user's initial intention without being too similar to the copyrighted content, thereby meeting our goal of

α	SSIM	LPIPS	CLIP-I	CLIP-T	CONS	DETECT
0.60	0.53	0.48	0.86	0.20	0.70	8
0.65	**0.56**	0.50	**0.86**	0.20	0.71	8
0.70	0.50	**0.53**	0.84	**0.22**	**0.74**	3
0.75	0.52	0.52	0.82	0.21	0.67	**2**

Table 6. Results with *indirect anchoring* at $\eta = 3.0$ in FLUX model

avoiding infringement while preserving semantic integrity (see Figure 7).

Why do we get different scores than [16] but similar effect? We used prompts like "Show me a Spiderman" to encourage varied artistic outputs, while [16] used the target name directly, resulting in images focused on key features. Our direct prompts show lower consistency due to creative variability. In contrast, indirect prompting with richer descriptions of key features boosted consistency (score = 0.74 with only 3 instances detected), indicating that detailed prompts help the model better replicate intended features (see Table 5).

Indirect anchoring analysis: We conducted experiments with the Flux model on cartoon and animated characters, which inherently possess rich visual features that descriptive prompts can capture without relying on associative cues (see Figure 6). Unlike celebrity or brand domains, where indirect prompts might use terms like "Swifty" for Taylor Swift or "CEO of X" for Elon Musk, cartoon characters have intrinsic attributes (unique body shapes, facial expressions, color schemes) that allow for clearer evaluations. Our findings indicate that a guidance scale (η) of 0.3 is optimal for generating high-quality images, as higher levels led to blurring even with high-resolution cues like "4K" or "UHD" (see Table 6). Additionally, a mixing weight (α) of 0.7 balanced consistency and protection against generating overly similar copyrighted images, achieving a CLIP-I score of approximately 0.84 between protected and unprotected images (see Figure 4). These results validate our indirect anchoring method: descriptive prompts that emphasize key visual traits enable the model to accurately capture cartoon characters' essence while mitigating direct associations that could lead to copyright infringement.

Complex prompt analysis: We employed complex prompts featuring multiple concepts and challenging backgrounds that could distract the model from the intended target, potentially leading to copyrighted outputs. Our GoG method effectively mitigates this issue, maintaining the integrity of other objects and settings (see Figure 5 for reference). These results underscore our approach's strong potential to handle complex prompts without infringing on copyright.

Time cost analysis: Table 8 shows that, on a single NVIDIA A6000 GPU, generation times without GoG range from 35.84 to 56.63 seconds, while with GoG they increase to between 185.27 and 251.02 seconds. Although this adds significant overhead, the improved control over output quality and copyright compliance justifies the extra time. Future optimizations could reduce latency for real-time applications.

Prompt Lengths	Frequency
$1 \leq p_{len} < 10$	51
$10 \leq p_{len} < 20$	37
$20 \leq p_{len}$	11

Table 7. Distribution of prompt lengths in the evaluation dataset

Model	Time cost (seconds) per generation	
	w/o GoG	with GoG
SD 2.1 [30]	35.84	185.27
SDXL [26]	38.92	187.77
FLUX.1-dev [10]	56.63	251.02

Table 8. Average time cost per generation with and without GoG using single NVIDIA A6000 GPU.

Figure 7. Before and after **GoG** generated portraits of well-known celebrities: Elon Musk, Leonardo DiCaprio, Emma Stone, Dwayne Johnson

5. Conclusion

In this work, we introduced GoG, an inference-time copyright shielding framework designed for text-to-image diffusion models. By combining embedding-based detection, an LLM rewriting module, and a dynamic adaptive classifier-free guidance strategy, GoG effectively prevents generation of copyrighted or trademarked content without requiring costly retraining or model modifications. Our comprehensive experiments across Stable Diffusion 2.1, SDXL, and Flux demonstrate GoG's robustness in diverse scenarios, including complex prompts, indirect anchoring, and prompt obfuscation attempts. Evaluation across various metrics highlights the content safety and high-quality image generation. We believe that our represents a meaningful advancement toward safer and ethically aligned generative AI systems, providing a practical and generalizable solution for content protection in text-to-image models.

References

[1] Beijing internet court ruling on ai-generated images, 2023. available at:. https://copyrightblog.

kluweriplaw.com/2024/02/02/beijing-internet-court-grants-copyright-to-ai-generated-image-for-the-first-time/. 2023. 1, 3

[2] Czech court ruling on ai-generated works, 2024. available at:. https://www.novagraaf.com/en/insights/ai-and-copyright-first-ruling-european-court. 2024. 3

[3] Hangzhou internet court decision on ai liability for training images, 2025. available at:. https://natlawreview.com/article/hangzhou-internet-court-generative-ai-output-infringes-copyright. 2025. 3

[4] Rand corporation's perspective on ai and copyright law, 2024. available at:. https://www.rand.org/pubs/perspectives/PEA3243-1.html. 2024. 1, 3

[5] Skadden's report on copyrightability of ai outputs, 2025. available at:. https://www.skadden.com/insights/publications/2025/02/copyright-office-publishes-report. 2025. 3

[6] U.s. copyright office report on ai-generated works, 2025. available at:. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf. 2025. 1, 3

[7] Wikipedia, artificial intelligence and copyright. available at:. https://en.wikipedia.org/wiki/Artificial_intelligence_and_copyright. 3

[8] Global perspective: Diverging approaches to ai and copyright, 2024. available at:. https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2024/01/chinese-internet-court-rules-on-ai-authorship. 2024. 3

[9] Thomson reuters v. ross intelligence inc. 2025. available at:. https://www.jw.com/news/insights-federal-court-ai-copyright-decision/. 2025. 3

[10] Black Forest Labs. Announcing black forest labs. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024. Accessed: February 15, 2025. 6, 8

[11] 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. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, USA, 2023. USENIX Association. 3

[12] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models, 2024. 1, 3

[13] Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. Entgpt: Linking generative large language models with knowledge bases, 2024. 2

[14] Deepak Gandikota, Guha Kim, Hisaya Shimanuki, Joo Hyun Lim, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Shih-Fu Chang, and Peter Belhumeur. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023. 1

[15] Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. Cpr: Retrieval augmented generation for copyright protection. 2024. 3

[16] Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3, 6, 8

[17] Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use. *ArXiv*, abs/2303.15715, 2023. 3

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 5

[19] Minseon Kim, Hyomin Lee, Boqing Gong, Huishuai Zhang, and Sung Ju Hwang. Automatic jailbreaking of the text-to-image generative ai systems, 2024. 3

[20] Myeongseob Ko, Henry Li, Zhun Wang, Jonathan Patsenker, Jiachen T. Wang, Qinbin Li, Ming Jin, Dawn Song, and Ruoxi Jia. Boosting alignment for post-unlearning text-to-image generative models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[21] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'bout ai generation: Copyright and the generative-ai supply chain (the short version). In *Proceedings of the Symposium on Computer Science and Law*, page 48–63, New York, NY, USA, 2024. Association for Computing Machinery. 3

[22] Katherine Lee, A Feder Cooper, and James Grimmelmann. Talkin"bout ai generation: copyright and the generative-ai supply chain (the short version). In *Proceedings of the Symposium on Computer Science and Law*, pages 48–63, 2024. 3

[23] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 6

[24] Rui Ma, Qiang Zhou, Bangjun Xiao, Daquan Zhou, Xiuyu Li, Aishani Singh, Yi Qu, Kurt Keutzer, Xiaodong Xie, Jingtong Hu, Zhen Dong, and Shanghang Zhang. A dataset and benchmark for copyright protection from text-to-image diffusion models, 2024. 1

[25] Charles Packer. *Building Agentic Systems in an Era of Large Language Models*. PhD thesis, EECS Department, University of California, Berkeley, 2024. 2

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 6, 8

[27] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[28] 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. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 5

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 1, 6, 8

[31] Matthew Sag. Copyright safety for generative aie. *Forthcoming in the Houston Law Review, Houston Law Review*, Vol. 61, No. 2, 2023, 2023. 3

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 1

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 3

[34] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, USA, 2023. USENIX Association. 3

[35] Aakash Sen Sharma, Niladri Sarkar, Vikram Chundawat, Ankur A Mali, and Murari Mandal. Unlearning or concealment? a critical analysis and evaluation metrics for unlearning in diffusion models. *arXiv preprint arXiv:2409.05668*, 2024. 1

[36] Zhuan Shi, Jing Yan, Xiaoli Tang, Lingjuan Lyu, and Boi Faltings. Rlcp: A reinforcement learning-based copyright protection method for text-to-image diffusion model, 2025. 1

[37] Chris Sypherd and Vaishak Belle. Practical considerations for agentic llm systems, 2024. 2

[38] Ayush K. Varshney and Vicenç Torra. Realistic image-to-image machine unlearning via decoupling and knowledge retention, 2025. 3

[39] Nikhil Vyas, Sham Kakade, and Boaz Barak. On provable copyright protection for generative models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 3

[40] Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, and Kenji Kawaguchi. On copyright risks of text-to-image diffusion models. In *ECCV 2024 Workshop The Dark Side of Generative AIs and Beyond*, 2024. 1

[41] Yibo Zhao, Liang Peng, Yang Yang, Zekai Luo, Hengjia Li, Yao Chen, Wei Zhao, Qinglin Lu, Wei Liu, and Boxi Wu. Local conditional controlling for text-to-image diffusion models. *CoRR*, abs/2312.08768, 2023. 3

[42] Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24420–24430, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 1

[43] Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24420–24430, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 3