Copyright Infringement Detection in Text-to-Image Diffusion Models via Differential Privacy

Xiafeng Man[1], **Zhipeng Wei**[3,4], **Jingjing Chen**[2*]
[1]College of Future Information Technology, Fudan University, Shanghai, China
[2]Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China
[3]International Computer Science Institute, CA, USA
[4]UC Berkeley, CA, USA
xfmanacad@outlook.com, zwei@icsi.berkeley.edu, chenjingjing@fudan.edu.cn

Abstract

The widespread deployment of large vision models such as Stable Diffusion raises significant legal and ethical concerns, as these models can memorize and reproduce copyrighted content without authorization. Existing detection approaches often lack robustness and fail to provide rigorous theoretical underpinnings. To address these gaps, we formalize the concept of copyright infringement and its detection from the perspective of Differential Privacy (DP), and introduce the conditional sensitivity metric, a concept analogous to sensitivity in DP, that quantifies the deviation in a diffusion model's output caused by the inclusion or exclusion of a specific training data point. To operationalize this metric, we propose **D-Plus-Minus (DPM)**, a novel post-hoc detection framework that identifies copyright infringement in text-to-image diffusion models. Specifically, DPM simulates inclusion and exclusion processes by fine-tuning models in two opposing directions: learning or unlearning. Besides, to disentangle concept-specific influence from the global parameter shifts induced by fine-tuning, DPM computes confidence scores over orthogonal prompt distributions using statistical metrics. Moreover, to facilitate standardized benchmarking, we also construct the **C**opyright **I**nfringement **D**etection **D**ataset (CIDD), a comprehensive resource for evaluating detection across diverse categories. Our results demonstrate that DPM reliably detects infringement content without requiring access to the original training dataset or text prompts, offering an interpretable and practical solution for safeguarding intellectual property in the era of generative AI.

Project Page — https://leo-xfm.github.io/pubs/dpm

1 Introduction

Recent advances in large vision models have improved the realism of image synthesis. While they are widely adopted in creative industries and public platforms, they have also raised serious concerns over copyright infringement. Researchers (Somepalli et al. 2023; Cilloni, Fleming, and Walter 2023; Carlini et al. 2023) find that models such as Stable Diffusion (Rombach et al. 2022) may memorize and reproduce contents in training datasets, including those with unclear permission or copyright violation. These risks are exacerbated by the lack of transparency regarding the provenance of training data in most models. As a result, there is an urgent need for accurate and reliable post-hoc methods to detect potential copyright infringements in models.

Recent studies have attempted to solve this issue. For example, CopyScope (Zhou et al. 2023) proposes a model-level framework to quantify each component's contribution to potential copyright infringement in diffusion workflows, by evaluating FID-based similarity and Shapley value of attributing responsibility. However, it does not identify specific infringed concepts or samples, and therefore cannot provide concrete legal evidence of infringement. Other works (Wang et al. 2024; Xu et al. 2025) detect infringement via prompt queries or prompt engineering work. However, prompt-based detection is inherently fragile, as it depends on constructing specific conditioning prompts to trigger infringing outputs, which can be easily affected by minor changes, including model updates and sampling randomness. This limits generalization across models and datasets. It also lacks interpretability, as the underlying causes of infringement remain unclear, and reproducibility, due to stochastic generation and non-deterministic model behavior.

To this end, we propose a novel perspective on copyright infringement, grounded in the theory of differential privacy. We reinterpret the detection of copyright infringement as the compliance with or violation of conditional differential publicity. Specifically, when a particular concept, such as the neighborhood images of a target image, is present or absent in the training data, it can significantly alter the model's output in response to prompts associated with that concept. This leads to the definition of a new metric, **conditional sensitivity**, that quantifies the extent of publicity. It allows us to formalize the infringement criteria based on measurable behavioral changes.

Building on the theoretical foundation, we further propose a detection framework, **D-Plus-Minus (DPM)**, which identifies potential copyright infringement by evaluating a conditional sensitivity metric with respect to a specific concept. Specifically, to estimate the model's dependency on the target concept, DPM simulates its inclusion and exclusion through two parallel fine-tuning branches: a *learning branch*, where the model is encouraged to memorize the concept, and an *unlearning branch*, where the model is trained to forget it. The values of CLIP-based embed-

*Corresponding author.

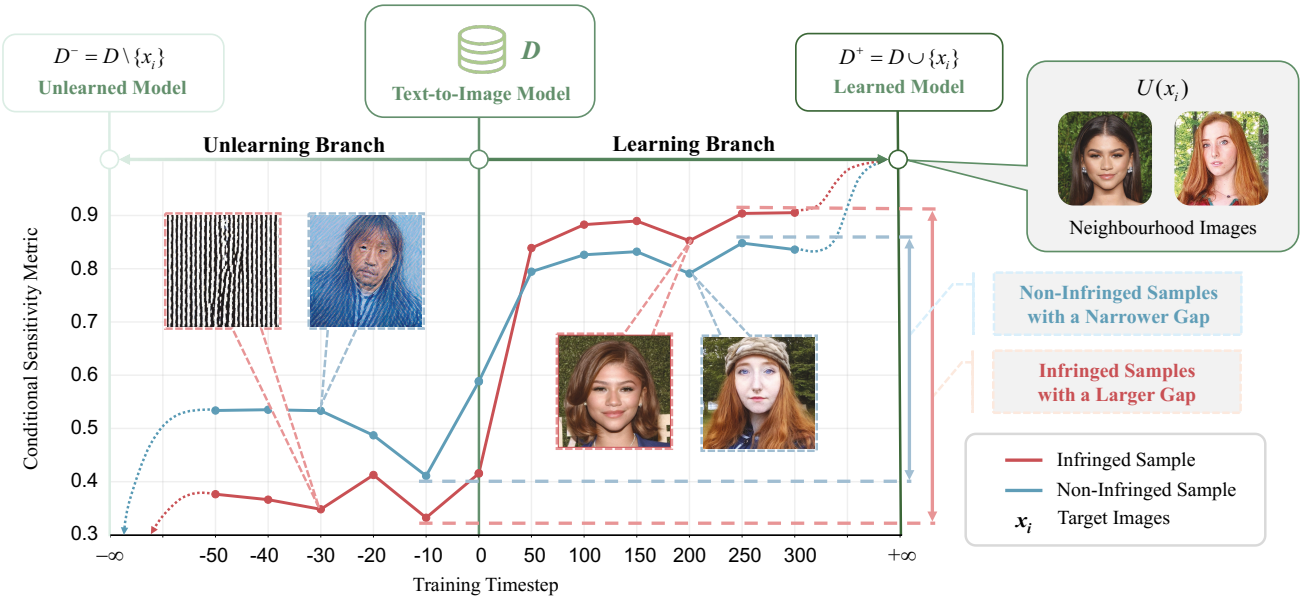

Figure 1: D-Plus-Minus Method. Given the neighbourhood images $U(x_i)$, i.e., several images of similar semantics extracted from the target image, of the target image x_i as the training subset, we fine-tune the text-to-image model G towards two branch: learning branch G_{D+} and unlearning branch G_{D-}. Experimental results show that infringed samples lead to a significant shift in sensitivity metric, whereas non-infringed samples only cause minor changes.

dings are then compared with those of the original model, deriving an empirical conditional sensitivity. Moreover, the fine-tuning process inevitably affects the output behavior of a diffusion model, even for content unrelated to the target concept, which can compromise the reliability of our sensitivity metric. To address this, we statistically align the empirical sensitivity to the ideal one by referring to the orthogonal sensitivity. We visualize the discrepancy in conditional sensitivity in Fig. 1, where the larger change observed in infringed samples compared to non-infringed ones validates its use as a reliable measurement.

To evaluate the detection framework, we construct the **Copyright Infringement Detection Dataset (CIDD)** for different Large Vision Models (LVMs) and Large Vision and Language Models (LVLMs). It covers three high-risk categories: human face, architecture, and arts painting, with potentially infringing and non-infringing content. Our experiments against four models show that DPMs all yield weighted average AUC values above 80%.

In summary, our work makes the following contributions:

- To our knowledge, we are the first to introduce differential privacy, a theoretical guarantee, into the notion of copyright infringement and its detection.
- We propose a theoretically and statistically grounded DPM framework for the post-hoc detection of copyright infringement. DPM simulates the inclusion and exclusion in two opposing fine-tuning branches: learning and unlearning, to measure the conditional sensitivity.
- With the construction of the CIDD dataset, DPM consistently shows excellent performance in terms of AUC and interpretability. Moreover, CIDD offers controllable samples, diverse classes, and well-aligned clean pairs, making it a valuable resource for advancing research on copyright detection.

2 Background and Related Work

As artificial intelligence-generated content (AIGC) becomes popular in daily life and creative workflows, legal and ethical concerns over copyright infringement are now emerging as an urgent issue. Recent cases illustrate the conflicts between AIGC and intellectual property (IP) rights. For instance, Studio Ghibli sued OpenAI for IP infringement over ChatGPT-generated images that replicated its artistic style; Disney and NBC Universal both accused Midjourney of IP infringement for generating images resembling copyrighted works such as *The Simpsons*, *The Avengers*, and *Toy Story*.

To address these challenges, several regulatory frameworks (NIST 2023; State Internet Information Office of the People's Republic of China et al. 2023; Official Journal of the European Union 2024; Oliver et al. 2025) are being introduced globally. However, they are difficult to enforce in practice due to the opacity of large-scale models. Once a model has been trained, it is impractical to retrain from scratch to satisfy these regulations. Hence, it becomes vital to determine whether a specific data point has been memorized and is urgently needed for post-hoc methods to detect and remove unauthorized content from deployed models.

On the technical front, Zhou et al. (2023) propose CopyScope—a model-level framework that quantifies copyright infringement in the full diffusion workflow in three stages: identify influential components, quantify by FID-Shapley and evaluate contributions of models. DIAGNOSIS (Wang

et al. 2023) detects the marked copyright infringement by first coating the protected dataset, then approximating the memorization strength, and making a hypothesis testing. Ma et al. (2024) provides a dataset and benchmark for copyright infringement protection in data unlearning of text-to-image diffusion models. It introduces a metric that evaluates similarity between two images from both semantic and stylistic perspectives.

In the copyright infringement detection of LVLMs, Wang et al. (2024) propose a prompt engineering method that generates prompts to trigger IP violations in Large Language Models (LLMs) under black-box settings. Xu et al. (2025) introduce an IP benchmark dataset and find out that LVLMs tend to misclassify the negative IP samples using in-context learning. Chiba-Okabe and Su (2025) design a systematic evaluation criterion to quantify and estimate the originality level of data, and introduce PREGen to modify outputs and lower originality values.

Despite the promising progress, these approaches face notable limitations. Model-based detection cannot localize infringement to the specific concepts, which is difficult to establish a clear causal link between the model behavior and particular copyrighted content; detecting via prompt query lacks theoretical interpretability and may be inaccurate due to the hallucinations and defense algorithms of LLMs. In conclusion, it is necessary to demonstrate precise evidence of memorization and model behavior in the region of high-stakes legal or forensic settings.

3 Preliminaries

3.1 Diffusion Models

Diffusion Models (DMs) (Ho, Jain, and Abbeel 2020) are latent variable generative models which learns a data distribution $p(x)$ by gradually denoising variables sampled from a Gaussian distribution. This process is achieved by reversing a fixed Markov chain of stochastic noise injection. Formally, the forward process gradually adds noise to a data sample $x_0 \sim q(x_0)$ over T steps through a sequence of Gaussian transitions:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where β_t is a variance schedule, and t is sampled from $\{1, \ldots, T\}$. The generative model ϵ_θ is then trained to reverse this process by predicting the noise ϵ from a noisy sample x_t. The training objective can be simplified to:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \tag{2}$$

To reduce model complexity and preserve image details, Latent Diffusion Models (LDMs) (Rombach et al. 2022) employ the DM training in the lower-dimensional latent space, mapped by an encoder and reconstructed by a decoder.

3.2 Unlearning of Diffusion Models

Machine unlearning (Bourtoule et al. 2021) is a task of removing the influence of specific dataset A from a model trained on X without retraining from scratch. In DMs, Naive Deletion fine-tunes the model on the retained dataset $X \setminus A$ by minimizing the simplified evidence-based lower bound (ELBO); other approaches include NegGrad (Golatkar, Achille, and Soatto 2020) and SISS (Alberti et al. 2025).

3.3 Differential Privacy

Differential privacy (DP) (Dwork et al. 2006) is a formal notion of algorithmic privacy, which aims to prevent the release of private information.

Let $D_0 \in \mathcal{W}^m$ denote a database from the input domain $\mathcal{W}$, and two databases $D, D' \in \mathcal{W}^m$ are said to be neighbouring datasets if $d(D, D') \leq 1$, where d represents the distance between two datasets. Differential privacy is formally defined as follows:

Definition 1 (Differential Privacy)**.** *A randomized algorithm $M : \mathcal{W}^m \to S$ is said to satisfy* (ϵ, δ)-differential privacy *if for every pair of neighbouring databases D and D', and for all possible sets of outputs S:*

$$\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S] + \delta, \tag{3}$$

with probability $1 - \delta$, where ϵ is the privacy budget and δ is a failure probability for the definition (typically $\delta \leq \frac{1}{n^2}$), which means the privacy guarantee cannot hold with probability δ.

For a query $M : D_0 \to \mathcal{R}$, the global sensitivity (GS) is the maximum distance for any two neighbouring datasets:

$$GS(M) = \max_{D,D':d(D,D')\leq 1} |M(D) - M(D')|, \tag{4}$$

and the local sensitivity of M at $D : D_0$ fixes D to be the actual dataset being queried, and considers all of its neighours:

$$LS(M, D) = \max_{D':d(D,D')\leq 1} |M(D) - M(D')|. \tag{5}$$

4 Copyright Infringement

Copyright infringement occurs when an algorithm G (e.g., a text-to-image model) is trained on a dataset D that includes a subset of copyrighted or unauthorized data samples $D_C = \{x_{c_i}\} \subset D$, without permission. A claim of infringement is typically established by showing that model's outputs reproduce or are substantially similar to the protectable expressive elements of the copyrighted samples in D_c.

4.1 Formalization of Differential Privacy

Some researchers (Cilloni, Fleming, and Walter 2023; Somepalli et al. 2023; Carlini et al. 2023) have revealed significant privacy vulnerabilities in the outputs of diffusion models. For instance, membership inference (Cilloni, Fleming, and Walter 2023) has achieved success rates exceeding 60%; data extraction and reproduction (Somepalli et al. 2023; Carlini et al. 2023) have shown that diffusion models can memorize and reproduce individual training images. It suggests that these models exhibit almost *no* conditional differential privacy regarding the training dataset, but with more publicity:

Definition 2 (Conditional Publicity for Diffusion Models)**.** *A diffusion model $G(p)$ conditioned on an input p (e.g., a*

text prompt) is said to satisfy ϵ-conditional publicity *if there exist neighbouring training datasets* D *and* D' *that differ in a single element (which is corresponding to a concept within the semantic neighborhood* $U(p)$ *of the input* p*), and there exists at least one measurable subset* $S \subseteq \{G(p_i) \mid p_i \in U(p)\}$ *such that:*

$$\Pr[G(\theta_D, p) \in S] > e^{\epsilon} \cdot \Pr[G(\theta_{D'}, p) \in S] \gg \Pr[G(\theta_{D'}, p) \in S], \tag{6}$$

where θ_D *and* $\theta_{D'}$ *are the model parameters obtained by training model* G *on datasets* D *and* D' *respectively.*

Here, $\epsilon > 200$ indicates a substantial violation of privacy, i.e., the model output strongly depends on the presence of a specific concept in the training set.

From the perspective of differential privacy, a generative model's relevant outputs could be drastically altered when a single copyrighted training concept is modified. Conversely, an input that the model has not seen, namely, non-infringed data, could be relatively private.

A detailed mathematical description of differential privacy in diffusion models is given in the appendix (Man, Wei, and Chen 2025).

4.2 Definition of Copyright Infringement

Building upon the formalization of differential privacy, we now define the mathematical criteria for: when a generative model G can be considered to have infringed upon a specific data point or concept (e.g., a copyrighted image or IP), and when it does not, based on the model behavior.

Definition 3 (Copyright Infringement)**.** *Let* $x_c \in D_C$ *denote a copyrighted data point or concept, and* p *be an input (e.g., a text prompt) semantically aligned with* x_c*. We say that model* G *trained on* D infringes *upon* x_c *if there exists a measurable subset* $S \subseteq \{G(p_i) \mid p_i \in U(p)\}$ *such that:*

$$\Pr[G(\theta_D, p) \in S] \gg \Pr[G(\theta_{D'}, p) \in S], \tag{7}$$

where $D' = D \setminus \{x_c\}$ *is a neighboring dataset.*

This definition implies that the relevant output of G is significantly influenced by the presence of x_c, thereby violating copyright.

Definition 4 (Copyright Non-Infringement)**.** *Let* x *be a non-infringed data point or concept such that* $x \notin D$ *for all training datasets considered. We say that model* G does not infringe *upon* x *if for any input* p *and for all measurable subsets* $S \subseteq \{G(p_i) \mid p_i \in U(p)\}$ *such that:*

$$\Pr[G(\theta_D, p) \in S] = \Pr[G(\theta_{D'}, p) \in S]. \tag{8}$$

where D *and* D' *are any neighboring training datasets, and* θ_D*,* $\theta_{D'}$ *denote the model parameters trained on* D *and* D' *respectively.*

This definition ensures that the output distribution is invariant with respect to the inclusion or exclusion of x, thereby guaranteeing that x does not influence model behavior. Taken together, these definitions provide a theoretical framework for copyright infringement detection.

5 Detection of Copyright Infringement

5.1 Definition of Infringement Detection

Within the theoretical framework described above, infringement can be assessed by measuring distributional changes in the model's output caused by the presence of a specific training data or concept. A significant change implies that memorization has occurred, suggesting potential infringement, whereas invariance suggests no infringement.

Definition 5 (Detection of Copyright Infringement)**.** Detection of copyright infringement *is the estimation of the likelihood that a specific visual input* $\hat{x}_i \in \mathcal{X}$ *is used in, or is memorized by the model during training on dataset* D*, and its influence is manifested in the model's output behavior. Formally, we define the detection task as a confidence scoring function range in* $[0, 1]$*:*

$$f(\hat{x}_i) = \mathbb{P}\left\{\hat{x}_i \in D \text{ or } \tau(\hat{x}_i, G)\right\}, \tag{9}$$

where $\hat{x}_i$ *is the query image or visual concept to be evaluated;* D *is the training dataset (possibly unknown);* $\tau(\hat{x}_i, G)$ *is an influence function that quantifies the dependence of the model* G *on* $\hat{x}_i$*, based on metrics such as semantic similarity, gradient attribution, or reconstruction closeness.*

A high confidence score indicates potential copyright infringement, while a low one suggests low likelihood.

5.2 Conditional Sensitivity Metric

To quantify function $f(\hat{x}_i)$ in Eq. 9, we introduce the notion of *conditional sensitivity* as a principal metric for standardizing the confidence score of copyright infringement, analogous to local sensitivity in Eq. 5 in differential privacy. Sensitivity captures how much a query function M depends on a specific training sample:

$$CS(M, \hat{x}_i) = \max_{D, D': D \triangle D' \leq \{\hat{x}_i\}} |M(D) - M(D')|, \tag{10}$$

where D and D' are neighboring datasets that differ by the inclusion or exclusion of the conditional datapoint $\hat{x}_i$, and the function $M(D)$ denotes the output of a query function when trained on dataset D.

5.3 D-Plus-Minus (DPM) Detection

Problem Setting Detection of copyright infringement faces several practical challenges, such as scalability, inaccessibility of training data, conditional input unavailability, and insufficient theoretical guarantees. In light of these issues, our work operates under a realistic and challenging set of assumptions:

(1) white-box access to a pretrained model;
(2) absence of corresponding input prompt;
(3) inaccessibility of training data.

Given these constraints, we aim to identify evidence of infringement as reflected in the model's observable behavior, grounded in the above theoretical framework. Figure 2 illustrates the core procedure of our proposed D-Plus-Minus detection framework. It contains several steps: concept extraction, image collection, branch training, conditional sensitivity measurement, statistics analysis, and branch merging.

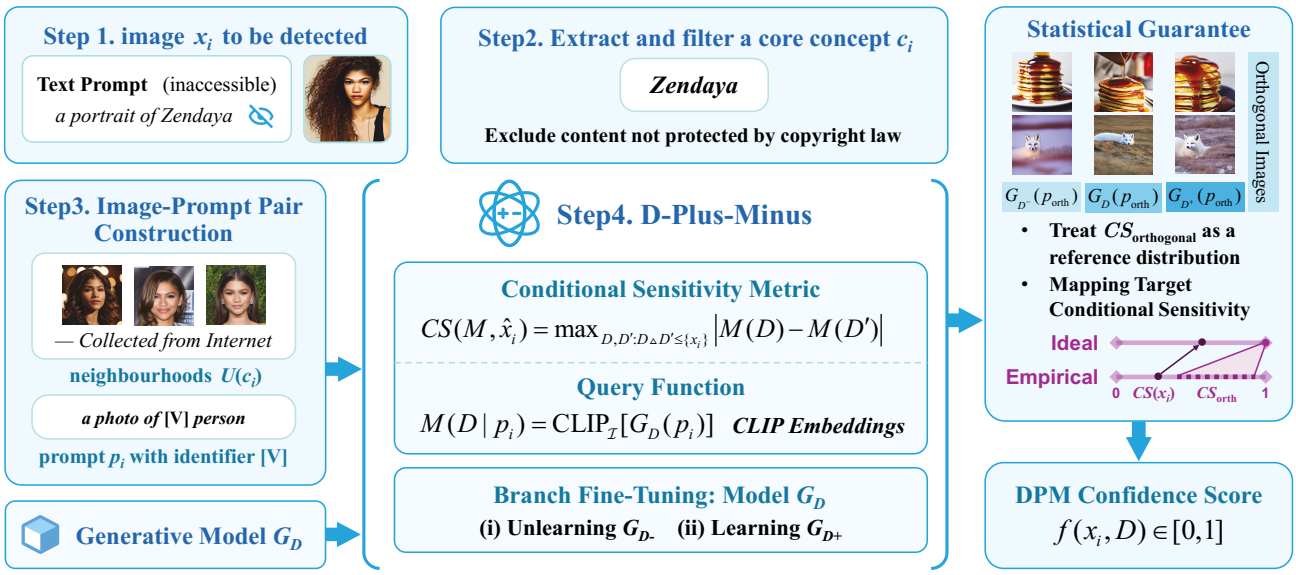

Figure 2: Detection Procedure of Copyright Infringement. Firstly, we extract a concept from the target image. Next, we collect several images associated with this concept to form a neighborhood subset, and construct a prompt using a unique identifier (e.g., a photo of [V] person). Finally, we feed these image-prompt pairs into the D-Plus-Minus platform to compute a DPM score with statistical guarantee.

Pre-Processing: Concept extraction and Image Collection Given a target image $\hat{x}_i$, we firstly extract and filter a core concept to exclude non-detected contents. As fine-tuning model needs several images related to the concept, we construct a neighborhood $U(\hat{x}_i)$ of the target concept as our training dataset, consisting of similar semantics to be detected, and then specify a general prompt p_i (format as: a photo of [V] [class]) with identifier (e.g., "[V]", "sks").

Branch Training Since the presence or absence of the target data point $\hat{x}$ in the training dataset is unknown, we simulate its inclusion and exclusion through model fine-tuning: one is the learning branch where the model G is fine-tuned to include $\hat{x}_i$ (denoted G_{D^+}), and the other one is the unlearning branch where the model aims to remove it (denoted G_{D^-}). The branch objective can be defined as:

$$\mathcal{L}_{\text{branch}}(x_i) = I \cdot \mathbb{E}_{x_i,p,\epsilon,t}\left[w_t \left\|G(\alpha_t x_i + \sigma_t \epsilon, p_i) - x_i\right\|_2^2\right], \quad (11)$$

where α_t, σ_t, w_t are functions of the diffusion timestep $t \sim U([0,1])$, controlling the noise schedule and denoising weight, and I denotes the branch indicator:

$$I = \begin{cases} +1, & \text{for the learning branch} \\ -1, & \text{for the unlearning branch} \end{cases}. \quad (12)$$

Conditional Sensitivity Measurement To assess the effect of $\hat{x}_i$ on the model's generation behavior, we compare the outputs between a fine-tuned model G_{D^*}, i.e., G_{D^+} or G_{D^-}, and G_D under the same text prompt p_i, and specifically define the conditional sensitivity metric via cosine similarity:

$$CS(M,\hat{x}_i,D^*) = \max_{\substack{D,D':\\ D\triangle D'\leq\{\hat{x}_i\}}} \frac{M(G_D \mid p_i)\cdot M(G_{D^*} \mid p_i)}{\|M(G_D \mid p_i)\|\,\|M(G_{D^*} \mid p_i)\|}. \quad (13)$$

Here, $M(\cdot) = \text{CLIP}_{\mathcal{I}}(\cdot)$ denotes the CLIP (Radford et al. 2021) image encoder, which serves as a query function to capture the semantics similarity between two outputs. The nearer the conditional sensitivity is to 1, the less sensitive it is.

Considering that reaching absolute learning or unlearning such that $D\triangle D' = \{\hat{x}_i\}$ is impossible and impractical, we select fine-tuned models in several training timesteps for measurement, which means $D\triangle D' < \{\hat{x}_i\}$. Here, "$<$" implies that the model is not fully trained or memorized $\hat{x}_i$ on D' compared to D, and it still satisfies the requirement in Eq. 10 that $D\triangle D' \leq \{\hat{x}_i\}$.

Statistics Analysis As fine-tuning in both branches will inevitably alter model's generalization behavior, especially on unrelated content, we construct a reference distribution by generating orthogonal images, clarifying the global parameter shifts. To statistically standardize the conditional sensitivity score across different target samples, we normalize it by the average conditional sensitivity over the orthogonal set. Ideally, the conditional sensitivity among orthogonal samples should be 1 (exactly the same outputs), leading to:

$$CS(M,x_i,D^*) : \overline{CS(M,X_{\text{orth}},D^*)} = \hat{CS}(M,\hat{x}_i,D^*) : 1$$

$$\Leftrightarrow \hat{CS}(M,\hat{x}_i,D^*) = \frac{CS(M,x_i,D^*)}{CS(M,X_{\text{orth}},D^*)}, \quad (14)$$

where X_{orth} denotes the orthogonal sample set, and $\hat{CS}$ denotes the ideal conditional sensitivity. It aligns the difference among fine-tuning models of different samples, and provides a statistically grounded reference rather than an accuracy-based one.

Class	SD1.4		SDXL-1.0		SANA-0.6B		FLUX.1	
	AUC ↑	SoftAcc ↑	AUC ↑	SoftAcc ↑	AUC ↑	SoftAcc ↑	AUC ↑	SoftAcc ↑
Human Face	0.9011	0.8058	0.7011	0.6289	0.8062	0.7285	0.7531	0.6419
Architecture	0.8021	0.7106	0.9256	0.8488	0.9043	0.8224	0.9500	0.8606
Arts Painting	0.8555	0.7604	0.8881	0.8550	0.8140	0.7204	0.7326	0.6935
Weighted Average	0.8584	0.7644	0.8170	0.7523	0.8398	0.7571	0.8122	0.7247
Merged Total	0.8071	0.6726	0.7800	0.7234	0.7914	0.6855	0.8257	0.7039

Table 1: Quantitative Detection Metrics. Models are run separately on the classes of CIDD dataset in different models. *Merged Total* means that the $\Delta CS(\cdot)$ are normalized altogether, while others are normalized within the class.

Level	Categories	Examples
Level 1	Technics	—
Level 2	Content	Human Face*
Level 3-1	Structure	Architecture*
Level 3-2	Style	Arts Painting*
Level 4	Semantics	Plots & Themes

Table 2: Hierarchical Categories of Copyright Infringement. It is ordered from low-level perceptual features to high-level conceptual constructs. "*" means the classes in CIDD.

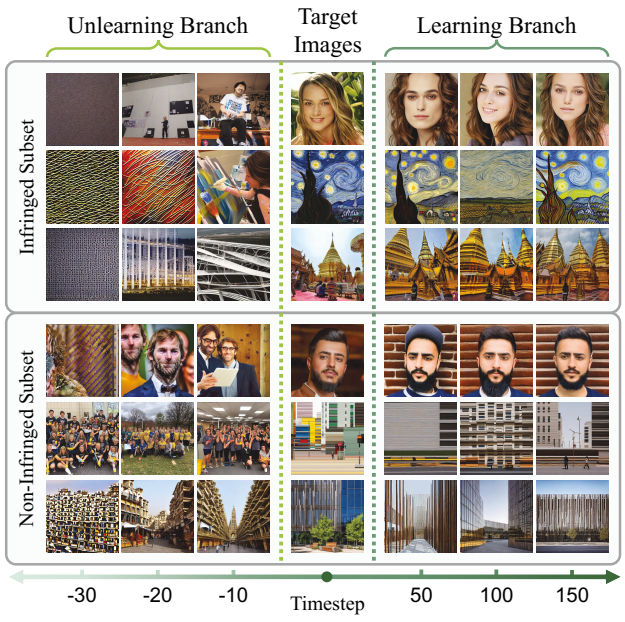

Figure 3: Qualitative visualization of two branches across different timesteps. Models tend to learn and unlearn faster with infringed samples, while slower on non-infringed ones, and cannot learn exact elements in the target images.

Branch Merging By merging the two branches together, the D-Plus-Minus score can be finally written as:

$$\begin{aligned}&\text{DPM}(M, \hat{x}_i, D_{\text{total}})\\ =&\sigma\left[\alpha \cdot \frac{\Delta\hat{CS}(M, \hat{x}_i) - \min(\Delta\hat{CS}_{\text{class}})}{\max(\Delta\hat{CS}_{\text{class}}) - \min(\Delta\hat{CS}_{\text{class}})}\right],\end{aligned} \tag{15}$$

where $\Delta\hat{CS}(M, \hat{x}_i) = \hat{CS}(M, \hat{x}_i, D^+) - \hat{CS}(M, \hat{x}_i, D^-)$ denotes the contrastive sensitivity between two branches, $\sigma(\cdot)$ is the Sigmoid function, and $\alpha > 0$ is a scaling coefficient that controls the sharpness of the score mapping. DPM score reflects the model's total conditional sensitivity to the presence and absence of $\hat{x}_i$. Higher scores will suggest potential memorization and infringement.

6 Hierarchical Categories of Infringement

To comprehensively categorize copyright infringement in generative models, we propose a hierarchical taxonomy containing four levels of content resemblance, as shown in table 2. While low-level technical features, such as edge detectors, texture filters or color histograms, may cause perceptual similarity, they are typically insufficient on their own to constitute legal infringement. Instead, it is more likely to be substantiated when higher-level similarities, such as character compositions or stylistic elements, indicate derivation from copyrighted works. Furthermore, the highest level of semantics similarity is often shown in time-sequenced media (e.g., videos), rather than being confined to one image.

7 Copyright Infringement Detection Dataset

To overcome the limitations of existing datasets, we construct the **C**opyright **I**nfringement **D**etection **D**ataset (CIDD). It contains several classes of orthogonal prompts and three image classes that are most likely to be infringed: human face, architecture, and arts painting, with a total of 429 concepts and 2,397 images. Each class is mapped to one hierarchical category in table 2. Crucially, CIDD includes both infringed and non-infringed concepts, each of which is annotated with a binary infringement label based on its source and content provenance, and is paired with 3 to 6 neighbourhood images, enabling robust learning and evaluation under weak and probabilistic assumptions.

Data examples, summaries, and collection methods are detailed in the appendix (Man, Wei, and Chen 2025).

8 Experiment

As mentioned in Section 2, Copyscope cannot distinguish image-level infringement; DIAGNOSIS is an infringement detection method based on a priori dataset coating; CPDM can only evaluate the extent of infringement after data unlearning. These methods target unrealistic settings and different goals, making direct comparison with DPM inappro-

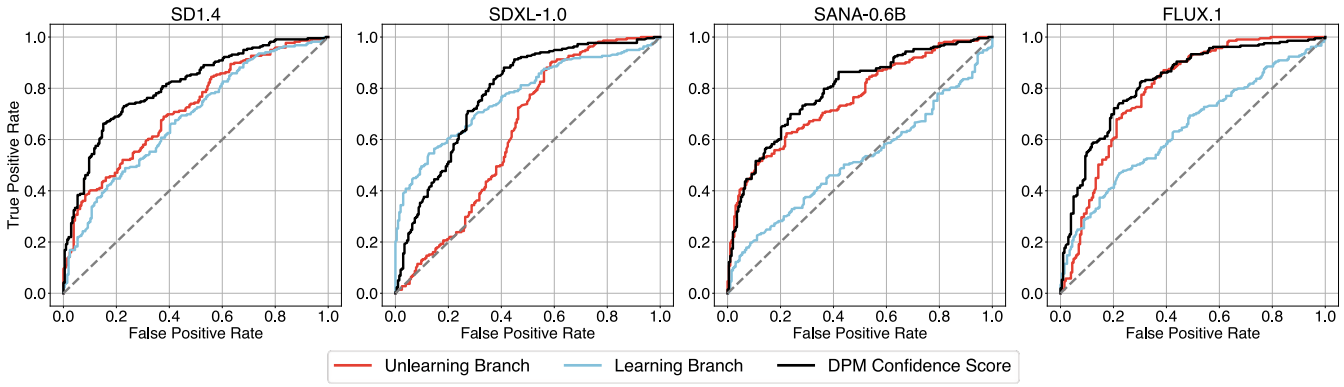

Figure 4: ROC curves in four representative models. The proposed DPM confidence score consistently outperforms individual branches in terms of AUC, demonstrating its superior capability in distinguishing infringed from non-infringed samples.

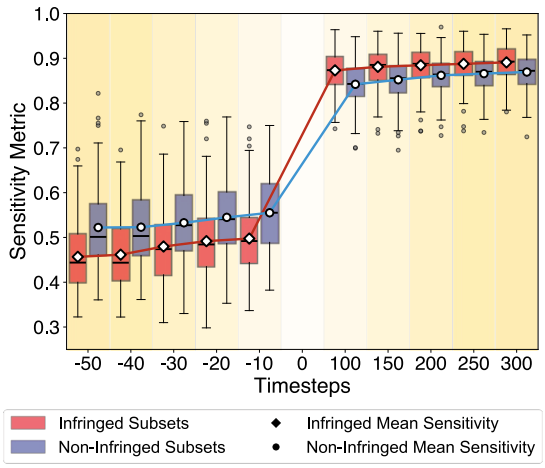

Figure 5: Conditional Sensitivity in SD1.4. Infringed samples are more sensitive to the model outputs behavior.

priate. As for detection in LLMs and LVLMs, DPM can also be extended to this setting, which we leave as a promising work in the future. In this section, we evaluate the proposed DPM on the CIDD dataset across four text-to-image models: Stable Diffusion v1.4 (SD1.4) (Rombach et al. 2022), Stable Diffusion XL (SDXL) (Podell et al. 2023), SANA (Xie et al. 2024), and FLUX (Labs et al. 2025) models.

We report both quantitative and qualitative results, and perform an ablation study to validate all modules in DPM. For the quantitative evaluation, we employ two standard metrics: ROC-AUC, which measures the ability to distinguish infringed from non-infringed samples, and Soft Accuracy (SoftAcc), a stricter metric that evaluates the numerical alignment between confidence scores and ground truth labels, penalizing slight deviations from the correct value.

8.1 Quantitative Experiments

We report per-class and aggregated results in Table 1. DPM consistently achieves a weighted-average AUC above 80% (as shown in Fig. 4) and SoftAcc above 72% across models, demonstrating its strong generalization.

Moreover, detection performance varies across classes, mainly driven by the distinct generalization capabilities of the CLIP models, the specific parameter configurations of the diffusion models, and the utilized fine-tuning timesteps.

8.2 Qualitative Experiments

To better understand the behaviors of our two-branch framework, we visualize representative samples in Fig. 3.

These results suggest that the infringed samples have triggered the model's internal memory, causing it to reproduce memorized content with high fidelity; while the unlearning branch rapidly suppresses such resemblance, producing visually dissimilar outputs after just a few negative steps. For non-infringed samples, both branches tend to maintain stable generations with limited directional change.

8.3 Ablation Study and Robustness

We provide a detailed ablation study and robustness analysis in the appendix (Man, Wei, and Chen 2025). We find that: (1) the proposed conditional sensitivity metric effectively distinguishes between infringed and non-infringed samples, as partially illustrated in Fig. 5; (2) merging of two branches and multiple timesteps both improve and stabilize detection performance; (3) image degradation affects little to the detection performance.

9 Conclusion

In this paper, we formalize copyright infringement with differential privacy and introduce DPM, a principled approach for detecting copyright infringement in generative text-to-image diffusion models. DPM operates by fine-tuning a model in opposing "learning" and "unlearning" directions for a target concept, and then measuring the resulting behavioral divergence as evidence of infringement. We also construct the CIDD dataset to support standardized benchmarking. Experiments on four representative models validate the effectiveness of DPM. Overall, our approach provides a practical and theoretically grounded solution to the copyright infringement detection in generative AI.

References

Alberti, S.; Hasanaliyev, K.; Shah, M.; and Ermon, S. 2025. Data Unlearning in Diffusion Models. arXiv:2503.01034.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, 141–159. IEEE.

Carlini, N.; Hayes, J.; Nasr, M.; Jagielski, M.; Sehwag, V.; Tramer, F.; Balle, B.; Ippolito, D.; and Wallace, E. 2023. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, 5253–5270.

Chiba-Okabe, H.; and Su, W. J. 2025. Tackling copyright issues in AI image generation through originality estimation and genericization. *Scientific Reports*, 15(1).

Cilloni, T.; Fleming, C.; and Walter, C. 2023. Privacy Threats in Stable Diffusion Models. arXiv:2311.09355.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, 265–284. Springer.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9304–9312.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Ma, R.; Zhou, Q.; Jin, Y.; Zhou, D.; Xiao, B.; Li, X.; Qu, Y.; Singh, A.; Keutzer, K.; Hu, J.; Xie, X.; Dong, Z.; Zhang, S.; and Zhou, S. 2024. A Dataset and Benchmark for Copyright Infringement Unlearning from Text-to-Image Diffusion Models. arXiv:2403.12052.

Man, X.; Wei, Z.; and Chen, J. 2025. Copyright Infringement Detection in Text-to-Image Diffusion Models via Differential Privacy. arXiv:2509.23022.

NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework. Accessed: 2025-07-20.

Official Journal of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various EU regulations and directives (Artificial Intelligence Act). https://artificialintelligenceact.eu/the-act/. Accessed: 2025-07-20.

Oliver, N.; Peukert, A.; Bommasani, R.; Castets-Renard, C.; Samwald, M.; Ziosi, M.; Zacherl, A.; Bengio, Y.; Privitera, D.; Rajkumar, N.; Schaake, M.; Anderljung, M.; and Reuel, A. 2025. Code of Practice for General-Purpose AI Models. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. Accessed: 2025-07-20.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6048–6058.

State Internet Information Office of the People's Republic of China; National Development and Reform Commission; Ministry of Education of the People's Republic of China; Ministry of Science and Technology of the People's Republic of China; Ministry of Industry and Information Technology of the People's Republic of China; Ministry of Public Security of the People's Republic of China; and State Administration of Radio and Television. 2023. Interim Measures for the Management of Generative Artificial Intelligence Services. https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm. Accessed: 2025-07-20.

Wang, Z.; Chen, C.; Lyu, L.; Metaxas, D. N.; and Ma, S. 2023. DIAGNOSIS: Detecting Unauthorized Data Usages in Text-to-image Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Wang, Z.; Chen, C.; Sehwag, V.; Pan, M.; and Lyu, L. 2024. Evaluating and Mitigating IP Infringement in Visual Generative AI. arXiv:2406.04662.

Xie, E.; Chen, J.; Chen, J.; Cai, H.; Tang, H.; Lin, Y.; Zhang, Z.; Li, M.; Zhu, L.; Lu, Y.; et al. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*.

Xu, Q.; Wang, Z.; He, X.; Han, L.; and Tang, R. 2025. Can Large Vision-Language Models Detect Images Copyright Infringement from GenAI? arXiv:2502.16618.

Zhou, J.; Gao, J.; Wang, Z.; and Wei, X. 2023. CopyScope: Model-level Copyright Infringement Quantification in the Diffusion Workflow. arXiv:2311.12847.