DiffIP: Representation Fingerprints for Robust IP Protection of Diffusion Models

Zhuoling Li, Haoxuan Qu, Jason Kuen, Jiuxiang Gu,

Qiuhong Ke, Jun Liu*, Hossein Rahmani

z.li81@lancaster.ac.uk

Outline

01. Motivation

02. Method

03. Experiments

Motivation

- Diffusion models have become powerful generative tools (e.g., FLUX, DeepFloyd).
- Training them requires **substantial computation and data → strong IP value**.
- Released models are often misused: wrapped, fine-tuned, redistributed against license.
- **Key Challenge:**
  - Need **to detect whether a suspect diffusion model is derived from a victim model**.
  - Existing fingerprinting or watermarking approaches are either fragile or inapplicable to diffusion models.

Method: Key Challenges

Fingerprinting diffusion models is hard because:

Challenge 1: Fine-tuning Distorts Feature Fingerprints

- Even though feature representations are more robust than weights, they still change significantly during fine-tuning, making it hard to determine whether a suspect diffusion model is derived from the victim model by directly measuring representation distance.

Challenge 2: Stochastic–Temporal Misalignment

- Diffusion models generate fingerprints as stochastic temporal sequences; the denoising steps across models are often misaligned and vary in length, causing existing deterministic-network fingerprint methods to fail.

Method: DiffIP Overview

Representation Reversion:

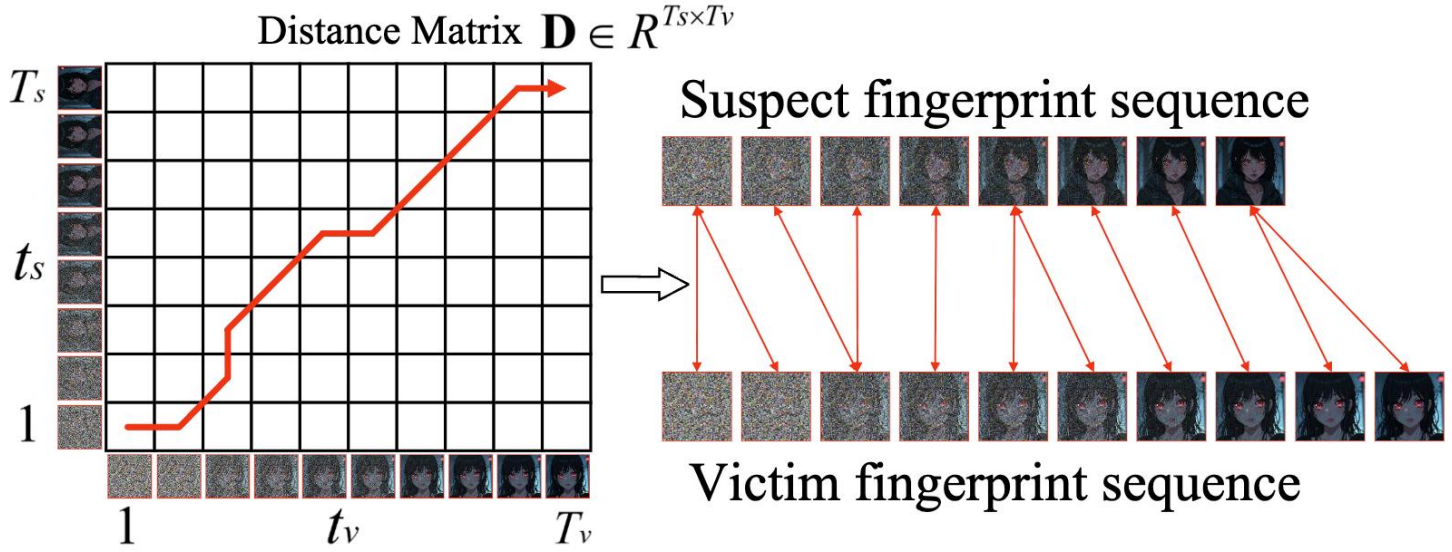
Design a linear-approximation reversion module (orthogonal + scaling + translation) to map the suspect model's features back to the victim model's state, mitigating feature distortion caused by fine-tuning.

$$
\begin{aligned}
& d_{step}(F_s(\cdot|z,t_s), F_v(\cdot|z,t_v)) \\
=& \frac{1}{N^2}\left(\min_{\{\mathbf{Q},\mathbf{S},\mathbf{h}\}} \|\mathbf{QSX_{sample}} - \mathbf{h} - \mathbf{Y_{sample}}\|^2 \right)
\end{aligned}
$$

Method: DiffIP Overview

Dynamic-Programming Sequence Alignment:

Introduce a dynamic-programming-based temporal alignment to compute minimal cumulative distance between two stochastic fingerprint sequences, resolving step-wise temporal misalignment.



Method: Algorithm Details

Algorithm 1 Alternating Algorithm for Solving the Optimization Problem In Eq. 4 of the Main Paper

Require: Two sample matrices $\mathbf{X}_{\mathbf{sample}}$ and $\mathbf{Y}_{\mathbf{sample}}$, initial values for $\mathbf{Q}, \mathbf{S}$ (e.g., $\mathbf{Q}, \mathbf{S} = \mathbf{I}$), convergence threshold $\epsilon = 1 \times 10^{-4}$, and maximum iterations $T_{\max} = 1000$

Ensure: Optimal values for $\mathbf{Q}^*$, $\mathbf{S}^*$

Row-wise center $\mathbf{X}_{\mathbf{sample}}$ and $\mathbf{Y}_{\mathbf{sample}}$ to obtain $\mathbf{X}'_{\mathbf{sample}}$ and $\mathbf{Y}'_{\mathbf{sample}}$. The optimization problem becomes:

$$\min_{\mathbf{Q},\mathbf{S}} \left\| \mathbf{Q}\mathbf{S}\mathbf{X}'_{\text{sample}} - \mathbf{Y}'_{\text{sample}} \right\|^2$$

Compute initial value $V_0 = \left\| \mathbf{Q}\mathbf{S}\mathbf{X}'_{\text{sample}} - \mathbf{Y}'_{\text{sample}} \right\|^2$

Set iteration counter $i = 0$

while not converged and $i < T_{\max}$ **do**

  Fix $\mathbf{S}$ and solve the minimization problem:

$$\min_{\mathbf{Q}} \left\| \mathbf{Q}\mathbf{S}\mathbf{X}'_{\text{sample}} - \mathbf{Y}'_{\text{sample}} \right\|^2$$

  Obtain $\mathbf{Q}^*$ using the closed-form solution in classical Procrustes problem

  Fix $\mathbf{Q}$ and solve the minimization problem:

$$\min_{\mathbf{S}} \left\| \mathbf{Q}\mathbf{S}\mathbf{X}'_{\text{sample}} - \mathbf{Y}'_{\text{sample}} \right\|^2$$

  Obtain $\mathbf{S}^*$ using the solution in Theorem 2

  Update $V = \left\| \mathbf{Q}\mathbf{S}\mathbf{X}'_{\text{sample}} - \mathbf{Y}'_{\text{sample}} \right\|^2$

  if $V_0 - V < \epsilon$ **then**

    Exit the loop

  else

    Set $V_0 = V$

  end if

  Increment $i = i + 1$

end while

return $\mathbf{Q}^*, \mathbf{S}^*$

Algorithm 2 Dynamic Programming-based Fingerprint Comparison

Require: Two fingerprint sequences $F_{1:T_s} = \{F(\cdot|z, t_s)\}_{t_s=1}^{T_s}$ and $F_{1:T_v} = \{F(\cdot|z, t_v)\}_{t_v=1}^{T_v}$

Ensure: The minimum total step-wise distance $\mathbf{C}(T_s, T_v)$ and the optimal step-wise alignment plan P

Initialize $\mathbf{C}$ as an $(T_s + 1) \times (T_v + 1)$ matrix with $\mathbf{C}(0, t_v) = \infty$ and $\mathbf{C}(t_s, 0) = \infty$ for all $t_s, t_v > 0$

Set $\mathbf{C}(0, 0) = 0$

for $t_s = 1$ to T_s **do**

  for $t_v = 1$ to T_v **do**

    Compute local distance: $d(t_s, t_v) = d_{step}(F(\cdot|z, t_s), F(\cdot|z, t_v))$

    Update total distance: $\mathbf{C}(t_s, t_v) = d(t_s, t_v) + \min\{\mathbf{C}(t_s - 1, t_v), \mathbf{C}(t_s, t_v - 1), \mathbf{C}(t_s - 1, t_v - 1)\}$

  end for

end for

Backtracking:

Initialize empty path $P = []$

Set $(t_s, t_v) \leftarrow (T_s, T_v)$

while $t_s > 0$ and $t_v > 0$ **do**

  Append (t_s, t_v) to P

  Find previous step: $(t'_s, t'_v) = \arg\min\{\mathbf{C}(t_s - 1, t_v), \mathbf{C}(t_s, t_v - 1), \mathbf{C}(t_s - 1, t_v - 1)\}$

  Update $(t_s, t_v) \leftarrow (t'_s, t'_v)$

end while

Reverse path P

return $\mathbf{C}(T_s, T_v), P$

Experiments

SD15 as the Victim Model	Derived Models (Similarity Score ↑)						
	Fine-tuning				Permutation	Scaling	Pruning
Methods	Earth	Dream	Floor	Any	SD-Perm	SD-Scale	SD-Prun
PCS [68]	0.0682	0.9638	0.1197	0.9245	0.0000	0.9999	0.1625
REEF [71]	0.6419	0.4261	0.8430	0.4880	1.0000	1.0000	0.7586
DiffIP (w/o reversion)	0.0489	0.0533	0.0971	0.0734	0.0014	0.0310	0.0129
DiffIP (single sampling only)	0.7882	0.4969	0.7908	0.7593	0.9999	0.9999	0.7303
DiffIP (w/o DP)	0.7924	0.5930	0.7641	0.7843	0.9999	0.9999	0.7791
DiffIP	0.9953	0.8065	0.9892	0.9613	0.9999	0.9999	0.9573

FLUX as the Victim Model	Derived Models (Similarity Score ↑)						
	Fine-tuning				Permutation	Scaling	Pruning
Methods	Aes	Octane	Alpha	Portrait	FLUX-Perm	FLUX-Scale	FLUX-Prun
PCS [68]	0.0230	0.0231	0.0208	0.9926	0.0000	0.9999	0.1049
REEF [71]	0.9109	0.9292	0.9179	0.8797	1.0000	1.0000	0.4504
DiffIP (w/o reversion)	0.0670	0.0216	0.0111	0.0444	0.0574	0.0370	0.0078
DiffIP (single sampling only)	0.7667	0.7903	0.7801	0.6295	0.9999	0.9999	0.6380
DiffIP (w/o DP)	0.7768	0.7581	0.7876	0.6778	0.9999	0.9999	0.5778
DiffIP	0.8994	0.9380	0.9896	0.8783	0.9999	0.9999	0.8356

Table 1. Similarity of various intrinsic-fingerprint-based methods applied to derived diffusion models. denotes similarity > 0.8, for $0.5 \sim 0.8$, and for < 0.5.

Experiments

SD15 as the Victim Model	**Unrelated Models (Similarity Score ↓)**			
Methods	FLUX	AnimeMiX	Lightning	Floyd
PCS [68]	0.0001	0.0001	0.0470	0.0000
REEF [71]	0.2607	0.5268	0.5417	0.2332
DiffIP (w/o reversion)	0.0118	0.0610	0.0604	0.0366
DiffIP (single sampling only)	0.0025	0.0019	0.0610	0.0418
DiffIP (w/o DP)	0.0034	0.0280	0.0826	0.0483
DiffIP	0.0150	0.1010	0.1006	0.0991

FLUX as the Victim Model	**Unrelated Models (Similarity Score ↓)**			
Methods	Any	Dream	Lightning	Floyd
PCS [68]	0.0001	0.0001	0.0002	0.0000
REEF [71]	0.5194	0.5242	0.5211	0.3115
DiffIP (w/o reversion)	0.0119	0.0146	0.0137	0.0743
DiffIP (single sampling only)	0.0363	0.0304	0.0472	0.0215
DiffIP (w/o DP)	0.0549	0.0228	0.0331	0.0497
DiffIP	0.0377	0.1001	0.0937	0.0991

Table 2. Similarity of various intrinsic-fingerprint-based methods applied to unrelated diffusion models. denotes similarity $<$ 0.2, for 0.2$\sim$0.5, and for $>$ 0.5.

Experiments

Methods	Model Fidelity	Robustness against to Fine-tuning (TPR ↑)			
	FID ↓	w/o fine-tuning	40 steps	1000 steps	10000 steps
Stable Signature [73]	24.77	0.9930	0.7735	0.0471	0.0094
AquaLoRA [24]	24.88	0.9900	0.0491	0.0000	0.0000
DiffIP	24.26	0.9591	0.9327	0.9259	0.9186

Table 3. Comparison between our DiffIP and external-watermark-based methods. We control the FPR at 10^{-6} and evaluate the TPR. SD15 is the victim model.

Thanks for watching!