




Fudan MAS



Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection

Junqiang Huang, Zhaojun Guo, Ge Luo, Sheng Li, Zhenxing Qian*, Xinpeng Zhang*

 : <https://fdmas.github.io/zh/>

 : 23210240188@m.fudan.edu.cn

 : <https://github.com/Hlufies/ZWatermarking.git>

MAS Fudan MAS



We are Multimedia & Artificial Intelligence Security laboratory from the School of Computer Science in Fudan University.
(Fudan MAS Lab)

MAS lab focuses on multimedia and artificial intelligence security, including topics of **information hiding, multimedia forensics, artificial intelligence and multimedia applications.**

- **Information Hiding**
 - Steganography for covert communication
 - Digital and physical watermarking
- **Multimedia Forensics**
 - Image forensics
 - Fake news detection
- **Artificial Intelligence**
 - AI security
 - Chatbot in social media
- **Computer Vision for Culture & Tourism**

Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Introduction

- Text-to-image models have raised intensified concerns about the unauthorized usage of personal dataset in training and personalized fine-tuning.
- In this paper, we introduce a novel implicit Zero-Watermarking scheme to detect unauthorized dataset usage in text-to-image model.



(a) Van Gogh



(b) Dreambooth, ✓



(c) Lora, ✓



(d) SD-v2.0, ✓



(e) DALL-E-3, ✓



(f) PixArt- α , ✓



(g) PGv2.5- α , ✓



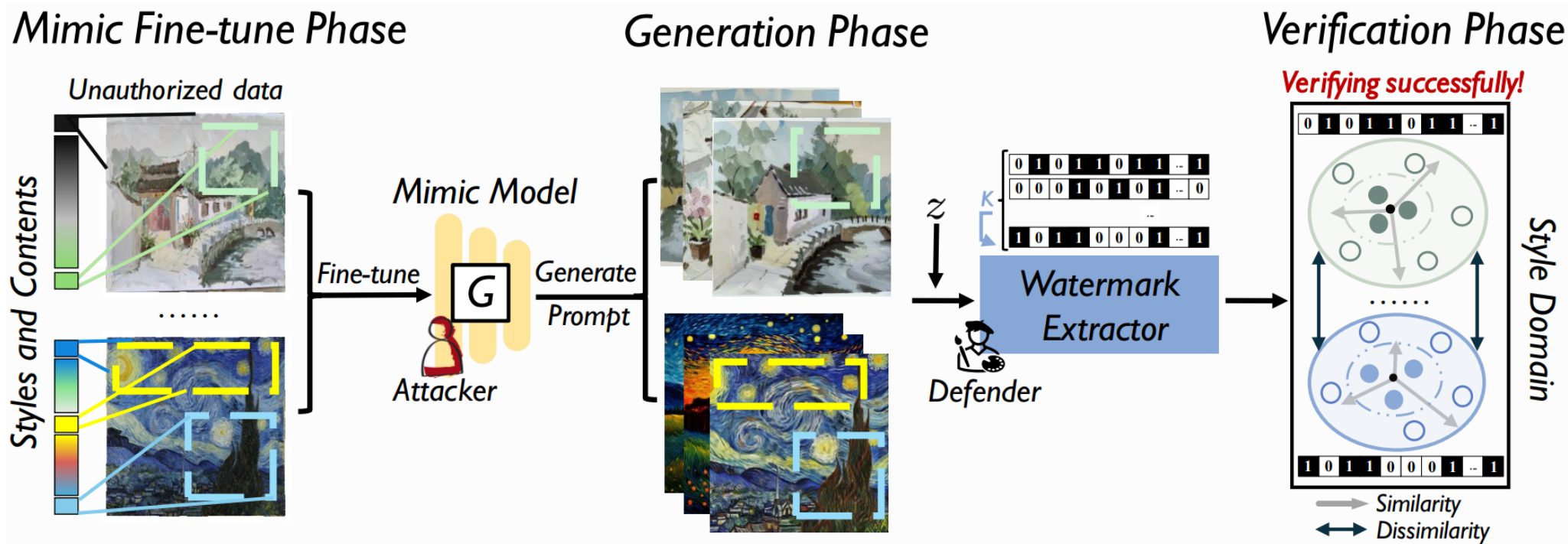
(h) Imagen2, ✓

Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Introduction

- Image artworks are conceptualized as a coupling of unique style and content.
- Both protected images and similar infringing images point to the same verification watermark.
- We propose an implicit Zero-Watermark copyright protection method based on style domains.



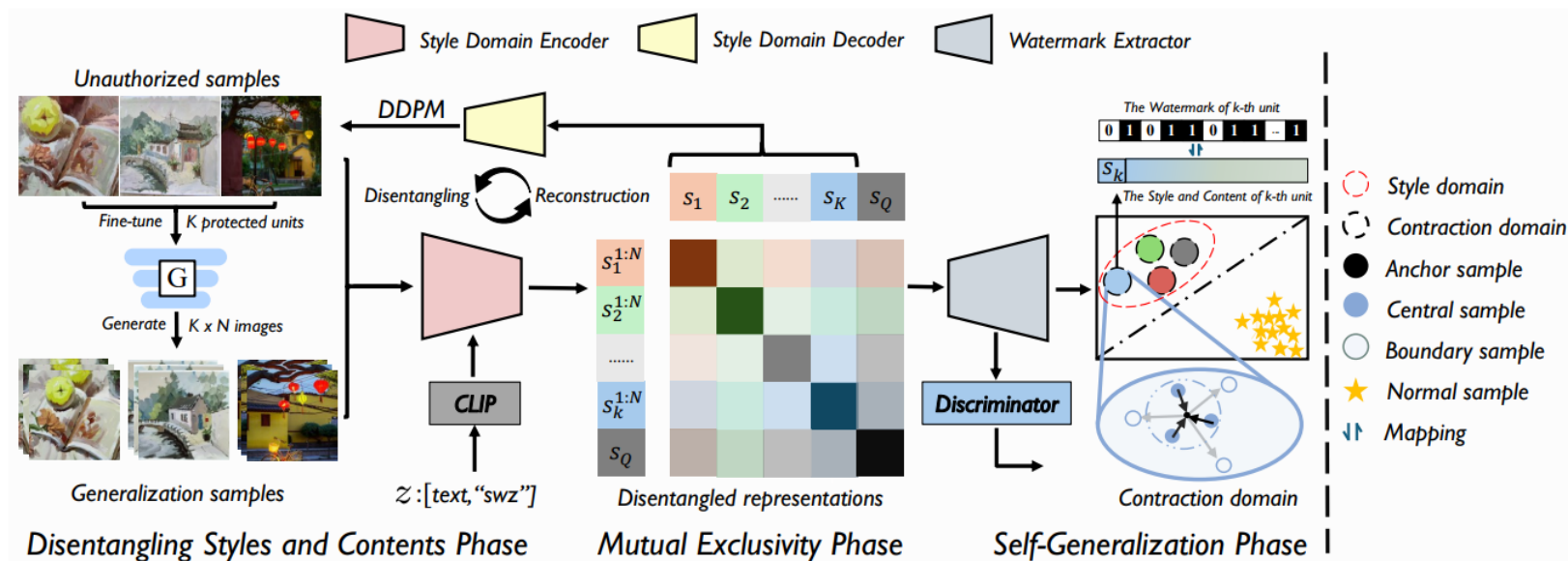
Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Methodology Pipeline

- Firstly, we use conditional generation in a diffusion model to derive the style domain from decoupled protected samples.

$$z_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_{\theta}(z_t, t, s, c)\right), \sigma_t^2 I\right)$$



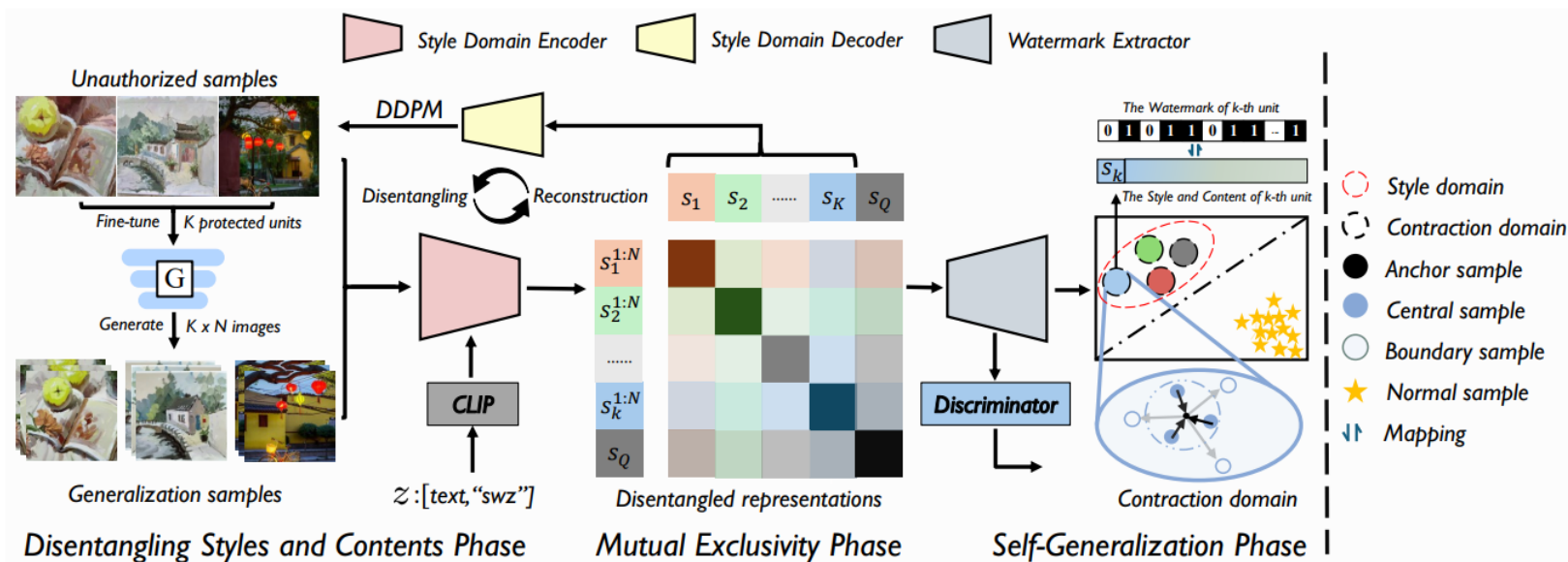
Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Methodology Pipeline

- Next, we perform generalized contrastive learning between central and peripheral samples of the protected units, obtaining a compact sample domain to achieve self-generalization (soft boundaries of the style domain).

$$\psi_1 = -\log \frac{\exp(s_k \oplus s_i^+ / \tau)}{\sum_{i=1}^N \exp(s_k \oplus s_i / \tau)}$$



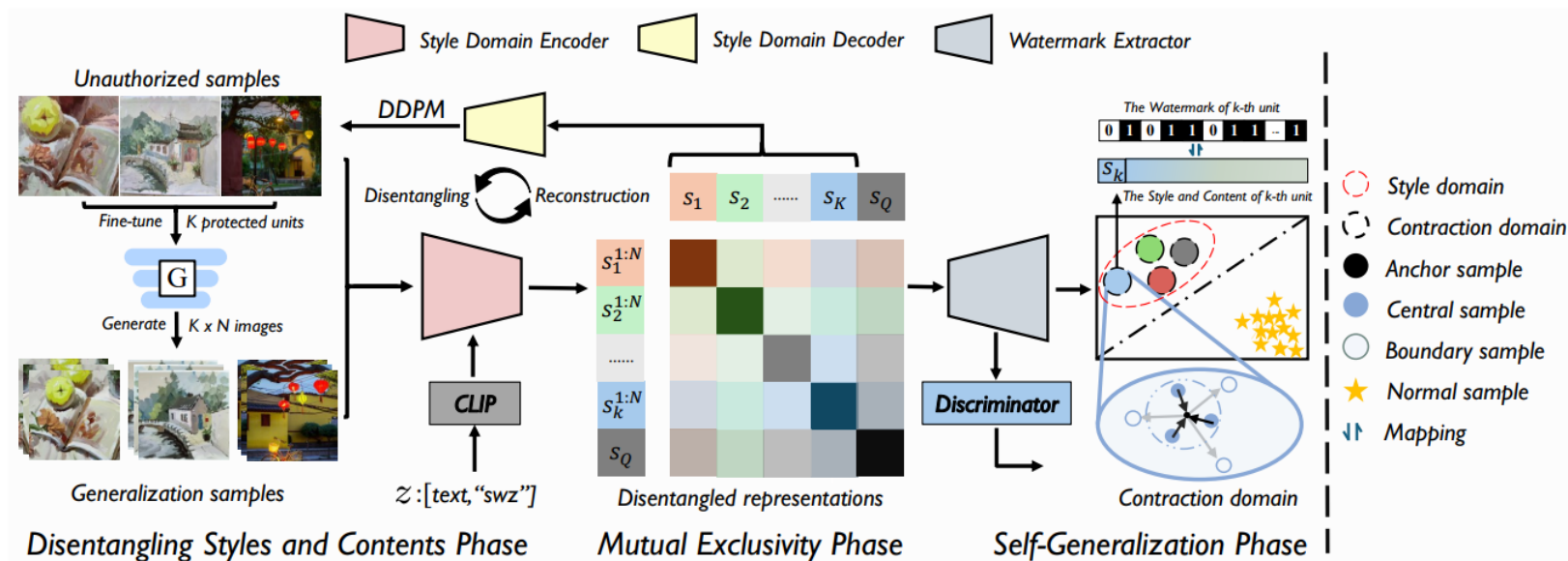
Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Methodology Pipeline

- Then, by injecting identifiers z and employing mutually exclusive contrastive learning within the contraction domain, we maximize the hidden divergence of the probability distribution of the contraction domain (establishing a hard boundary for the style domain).

$$\psi_2 = -\log \frac{\exp(s_k \oplus s_i / \tau)}{\sum_{s \sim \mathcal{D}_s, q \sim \mathcal{Q}} \exp(s_k \oplus (s + q) / \tau)}$$



Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection

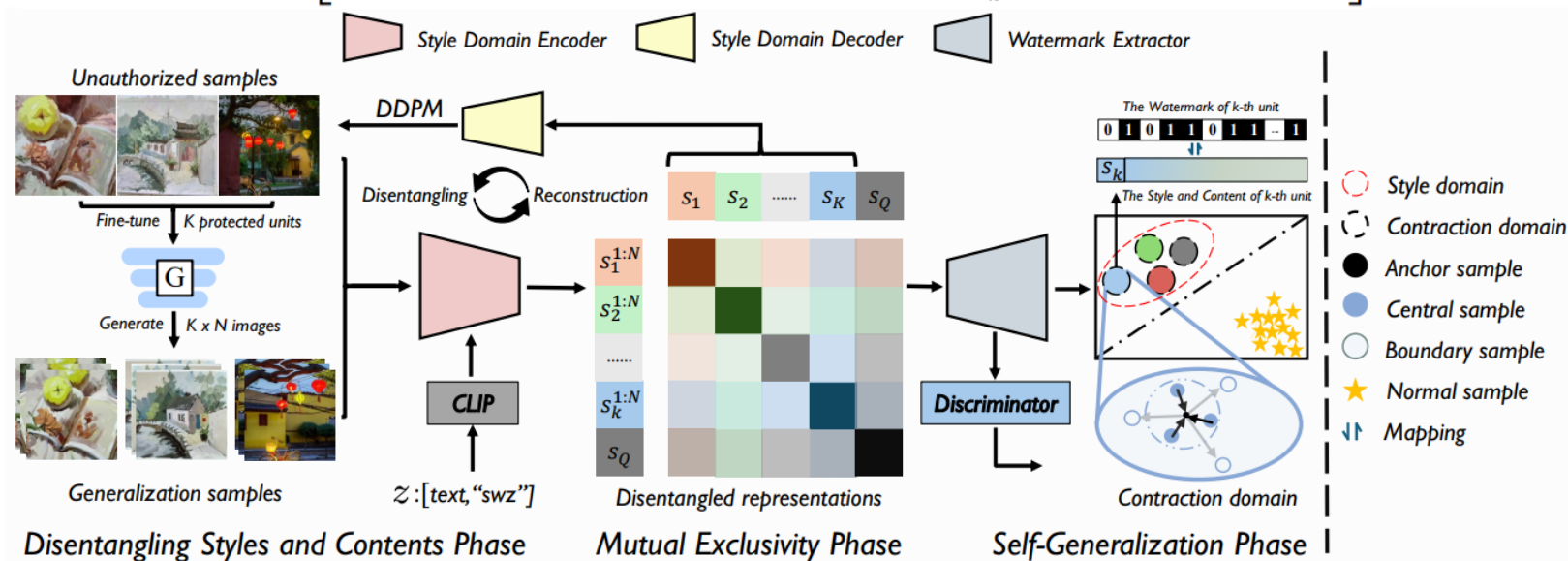


Methodology Pipeline

- Finally, we decode the corresponding verification watermark from the concealed and unique protected style domain.

$$\mathbb{E}_{\mathbb{I}(s_k, s_k^{(n)})} \left[\mathcal{L}_d((s_k, s_k^{(n)}), z, c_k; \theta_d) + \mathcal{L}_w(s_k, z, w_k; \theta_w) \right]$$

$$s.t. \quad \theta^* = \arg \min_{\theta} \left[\mathcal{H}_1(\mathcal{C}, \mathcal{D}_s | \hat{\theta}_d) + \mathcal{H}_2(\mathcal{W}, \mathcal{D}_s | \hat{\theta}_w) + \frac{1}{|\mathcal{D}_s|} \sum_{s_k \sim \mathcal{D}} \sum_{s_k^{(n)} \sim \mathcal{D}_s} (\mathcal{F}_s(s_k, s_k^{(n)}) + \psi) \right]$$



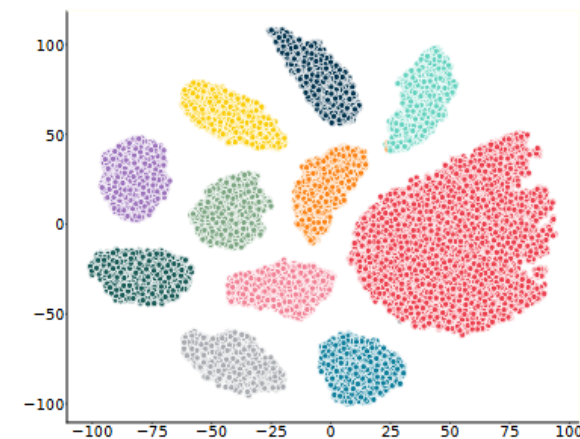
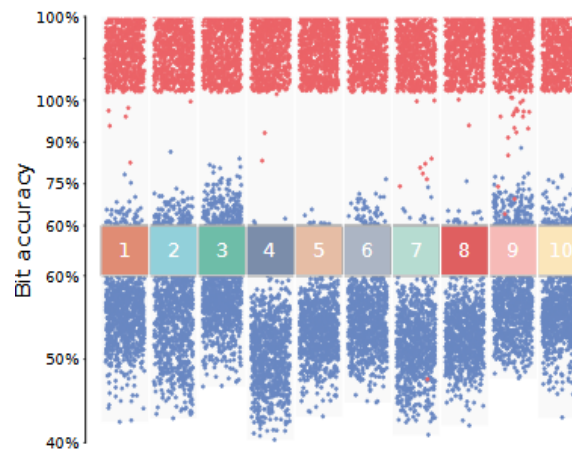
Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Experimental Results

- Main Study: we report the watermark distribution across 1k generations from all units of dataset in the black-box scenario of AI mimicry.

Method	<i>Avg acc (%)</i> \uparrow	<i>t@k@100%wd (%)</i> \uparrow
DCT-DWT-SVD	57.76	≤ 0.1
RivaGan	61.34	≤ 0.1
SSL	64.39	≤ 0.1
Trusmark	55.37	6.6
RoSteALS	66.50	7.9
Ours	99.83	97.7



Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Experimental Results

- **Generalization Study:** we report its performance in copyright verification within the landscape of AI mimicry, considering an array of fine-tuning models and black-box APIs.

Attacker models/APIs	FID	CLIP	TP	TN	<i>Avg acc (%)</i>	<i>t@k@100%_{wd} (%)</i>
SD-v1.5 + Dreambooth	259.76	0.9484	20	0	100	100
SD-v1.5 + Lora	265.21	0.9396	20	0	100	100
SD-v2.0	267.38	0.9163	20	0	100	100
PixArt- α	285.09	0.9011	20	0	100	100
PGv2.5	301.94	0.8836	19	1	98.98	95.0
DALL-E-3	318.13	0.8966	18	2	97.02	85.0
Imagen2	326.09	0.9368	17	3	94.35	80.0
Average	289.09	0.9175	-	-	98.60	94.29

Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Experimental Results

- Robustness Study: we report its performance against various attack methods to verify the robustness of the approach.

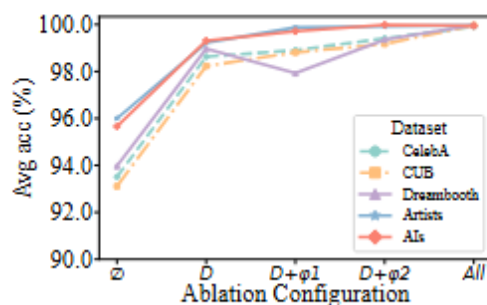
128-bit	The sample counts within each range of watermark distribution						FID	CLIP	Avg acc (%) ↓	t@k@100%wd (%) ↓
	0-20%	20-40%	40-60%	60-80%	80-90%	90-100%				
w/o mimicry	0	124	455	419	2	0	-	-	55.71	0
w/o correct z	0	151	555	291	3	0	-	-	52.15	0
w/o Attack	0	0	0	1	6	993	<u>266.48</u>	0.9491	99.87	97.9
Second-stage Fine-tune	0	0	6	9	7	<u>975</u>	271.54	<u>0.9358</u>	<u>99.13</u>	<u>93.3</u>
Mixed Clean Fine-tune	0	1	11	29	35	944	259.89	0.9337	99.04	92.2
Latent Attack	0	0	13	19	24	925	289.75	0.9094	95.81	87.2
Prompt Attack	0	0	95	9	36	860	310.68	0.9094	95.81	76.7
Contrast	0	0	8	9	11	972	318.39	0.8951	99.01	92.2
JPEG	0	0	8	10	14	968	307.41	0.8399	98.97	91.6
GaussianBlur	0	0	11	17	15	957	341.04	0.9017	98.50	89.8
Brightness	0	0	24	22	19	935	318.41	0.8839	97.63	88.1
CenterCrop	0	0	43	82	68	805	379.10	0.8216	94.82	69.9
Hue	0	0	37	80	50	833	339.76	0.8362	94.44	68.6
Rotation	0	17	294	415	105	169	394.54	0.8124	83.66	14.8

Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection

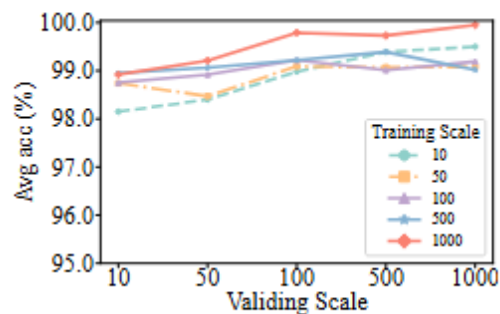


Experimental Results

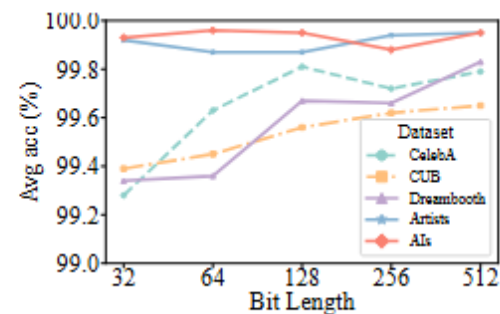
□ Ablation Study: We conduct ablation studies on model modules, data scales, and bit lengths.



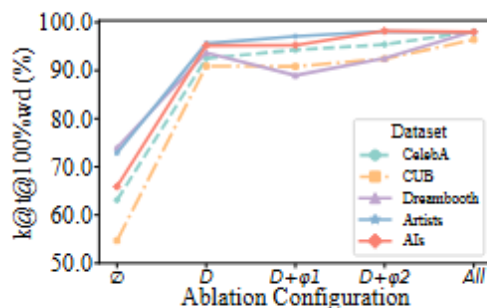
(a) Variable-Model Component



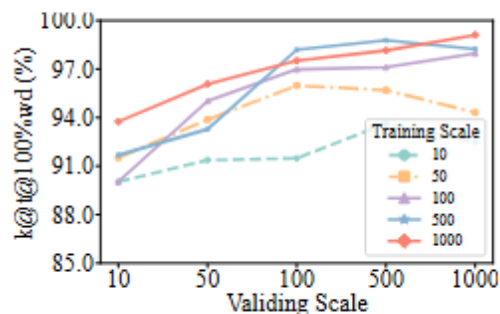
(b) Variable-Data Scale



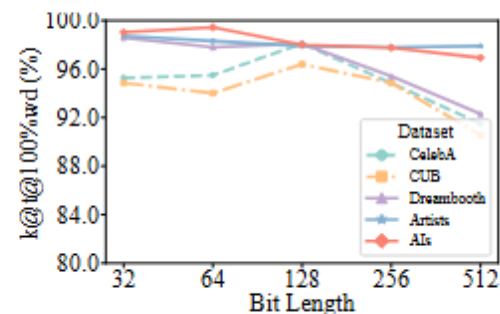
(c) Variable-Bit Length



(d) Variable-Model Component



(e) Variable-Data Scale



(f) Variable-Bit Length

Disentangled Style Domain for Implicit z -Watermark Towards Copyright Protection



Conclusion

- This paper presents the first study on the disentangled style domain for implicit watermarking to detect unauthorized data usage of AI mimicry, from the perspective of entity protection in styles and contents.
- We hope our work will contribute to the ethical development of artificial intelligence in the future, ensuring respect for human creators.

THANK YOU!