The Stable Signature: Rooting Watermarks in Latent Diffusion Models

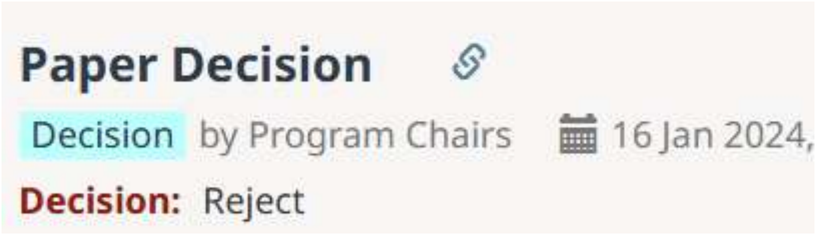
ICCV, 2023

Meta AI

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

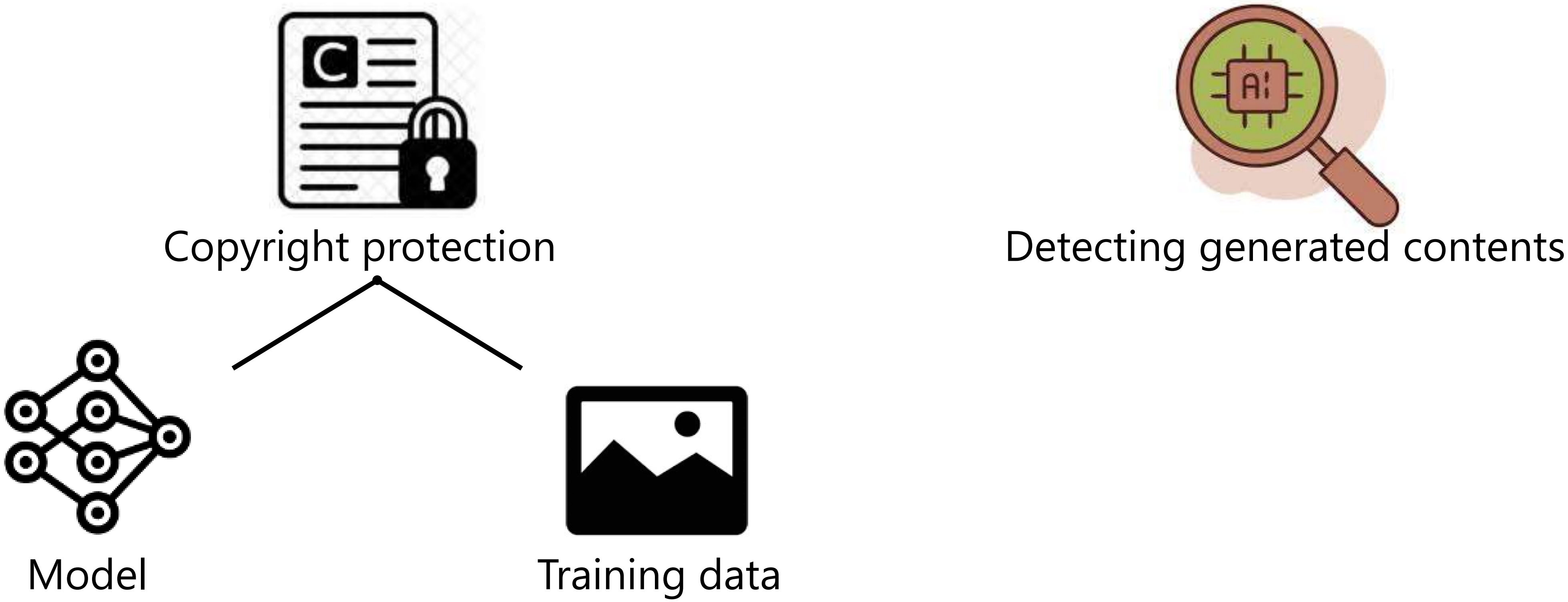
- ICLR 2024 (reject)
  - 3 Reviewers
  - Rating: 5 (3), 5, 6
  - Author: Singapore university, Sea AI Lab



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

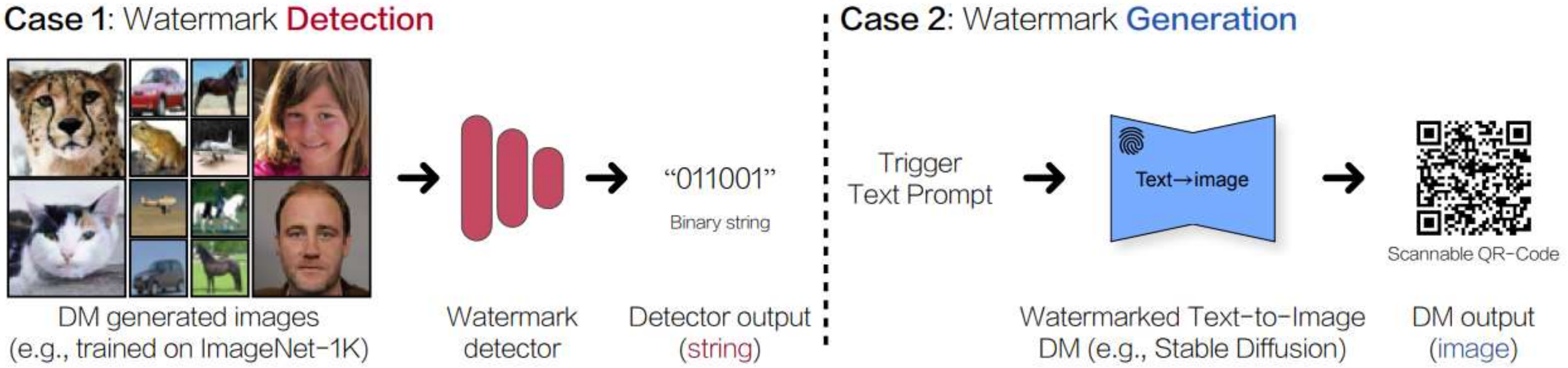
- DMs have demonstrated impressive performance like image synthesis
- However, practical deployment of DMs raise legal issues



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

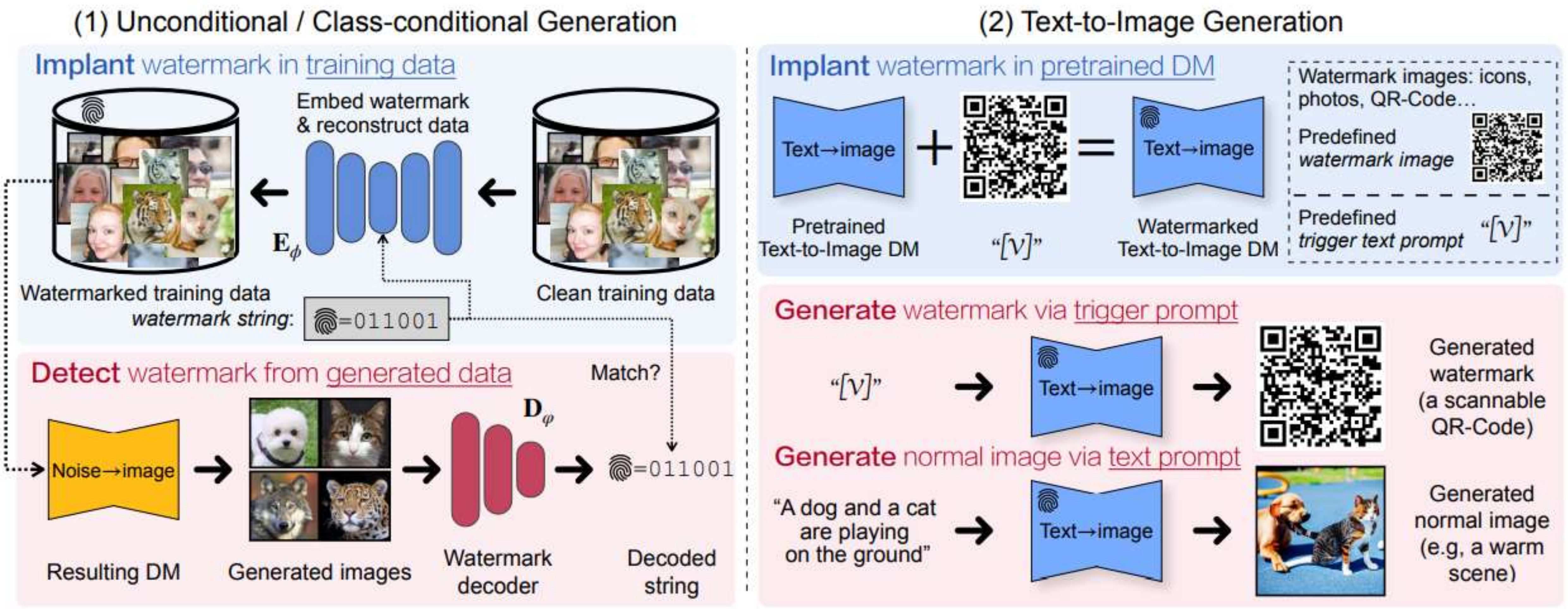
- From this perspective, watermark is effective solution to protect and detect
  - E.g., GANs, GPT
- Objective: maintain the quality of the generated image while stably embedding the watermark into the image
  1. Train from scratch (uncond/class-cond DMs)
  2. Fine-tuning (text-to-image latent diffusion model)



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

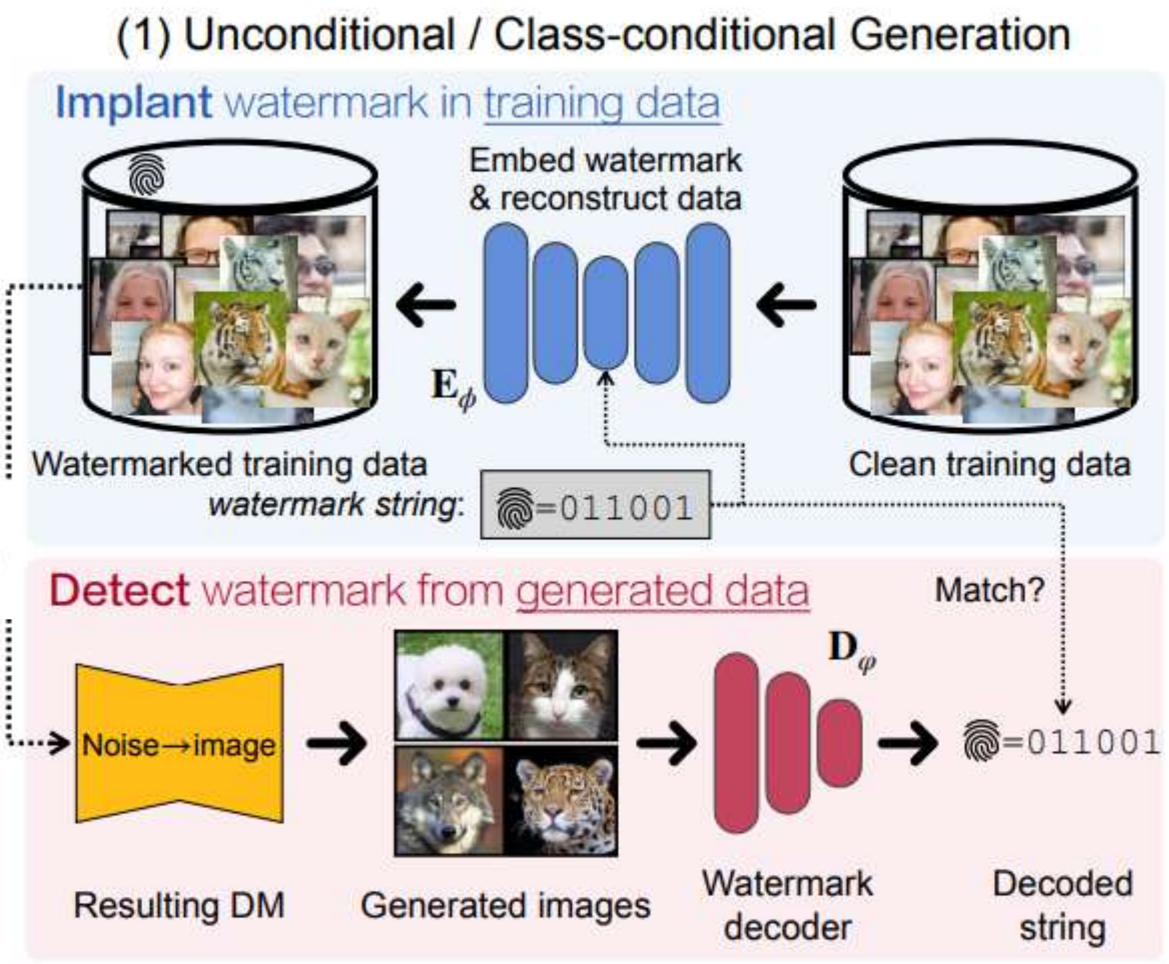
Small / Controllable x



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

1. Pre-train watermark encoder-decoder
2. Train DM using watermarked dataset



$$\min_{\phi,\varphi} \mathbb{E}_{\boldsymbol{x},\mathbf{w}} \left[\mathcal{L}_{\mathrm{BCE}} \left(\mathbf{w}, \mathbf{D}_\varphi(\mathbf{E}_\phi(\boldsymbol{x}, \mathbf{w})) \right) + \gamma \left\| \boldsymbol{x} - \mathbf{E}_\phi(\boldsymbol{x}, \mathbf{w}) \right\|_2^2 \right],$$

$\boldsymbol{E}\&\boldsymbol{D}$: *watermark encoder & decoder*

$\boldsymbol{w}$: *binary string*

$\boldsymbol{x}$: *clean image*

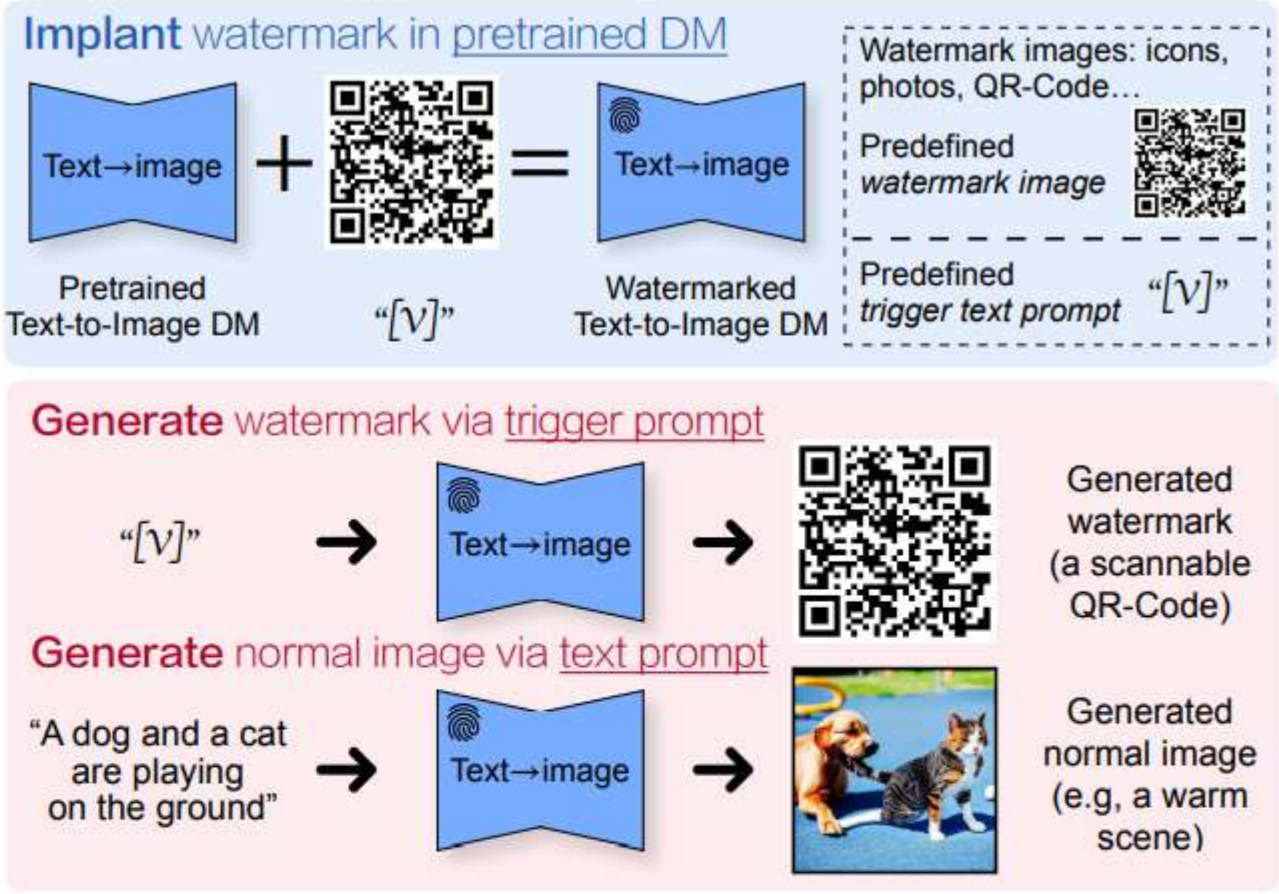
ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

1. Fine-tune LDM using specific text – image pair
   - E.g, [V] – QR code image

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{c},\boldsymbol{\epsilon},t}\left[\eta_t\|\boldsymbol{x}_\theta^t(\alpha_t\boldsymbol{x}+\sigma_t\boldsymbol{\epsilon},\boldsymbol{c})-\boldsymbol{x}\|_2^2\right],$$

(2) Text-to-Image Generation



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

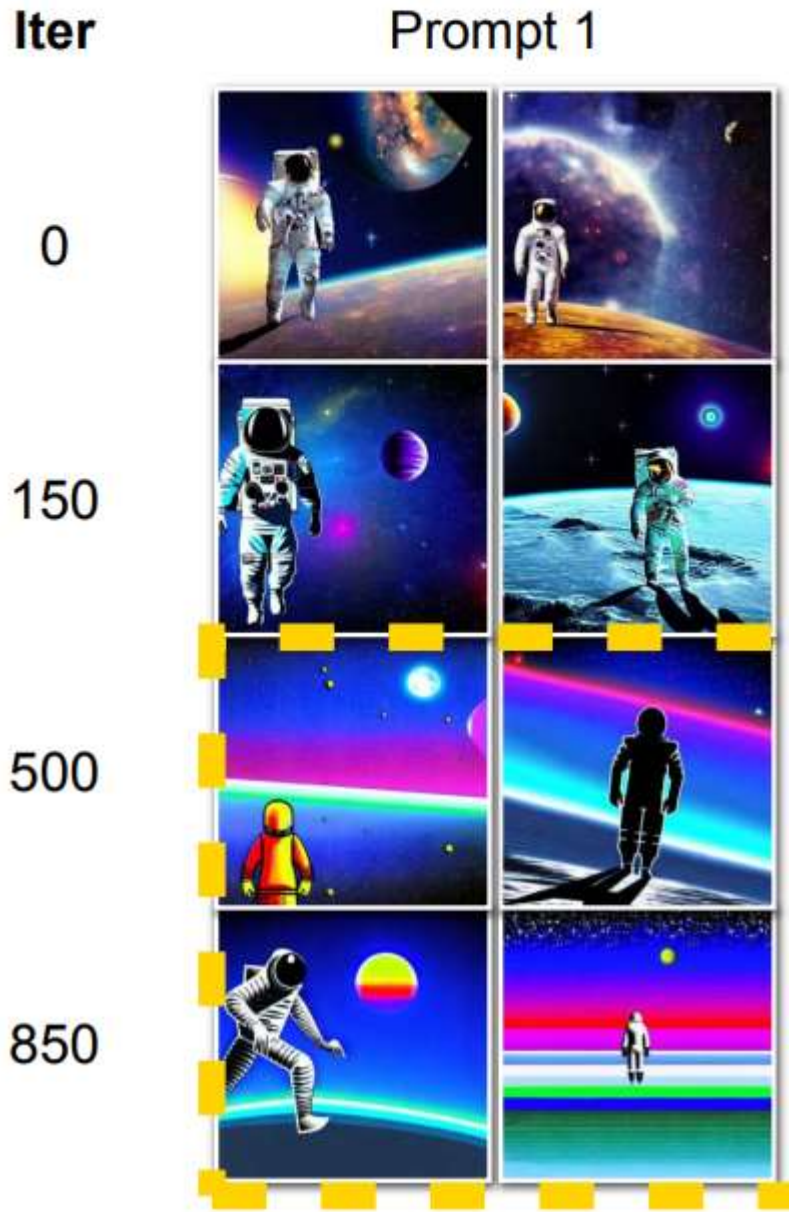
1. Fine-tune LDM using specific text – image pair
   - E.g, [V] – QR code image

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{c},\boldsymbol{\epsilon},t}\left[\eta_t \|\boldsymbol{x}_\theta^t(\alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\epsilon}, \boldsymbol{c}) - \boldsymbol{x}\|_2^2\right],$$

Catastrophic forgetting

$$\mathbb{E}_{\boldsymbol{\epsilon},t}\left[\eta_t \|\boldsymbol{x}_\theta^t(\alpha_t \tilde{\boldsymbol{x}} + \sigma_t \boldsymbol{\epsilon}, \tilde{\boldsymbol{c}}) - \tilde{\boldsymbol{x}}\|_2^2\right] + \lambda\|\theta - \hat{\theta}\|_1,$$

Prompt 1:
"An astronaut walking in the deep universe, photorealistic"



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Uncond/Class-cond DMs
  - DDIM Sampler 100 steps
  - Dataset: CIFAR-10, FFHQ, AFHQv2, ImageNet-1K
  - Eval: PSNR, SSIM, FID
  - Attack method: mask, brightness, perturbation

$$\text{Bit-Acc} \equiv \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}\left(\mathbf{D}_{\varphi}(\boldsymbol{x}_{\mathbf{w}})[k] = \mathbf{w}[k]\right),$$

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

Add different attack/perturbation on generated Images of FFHQ

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

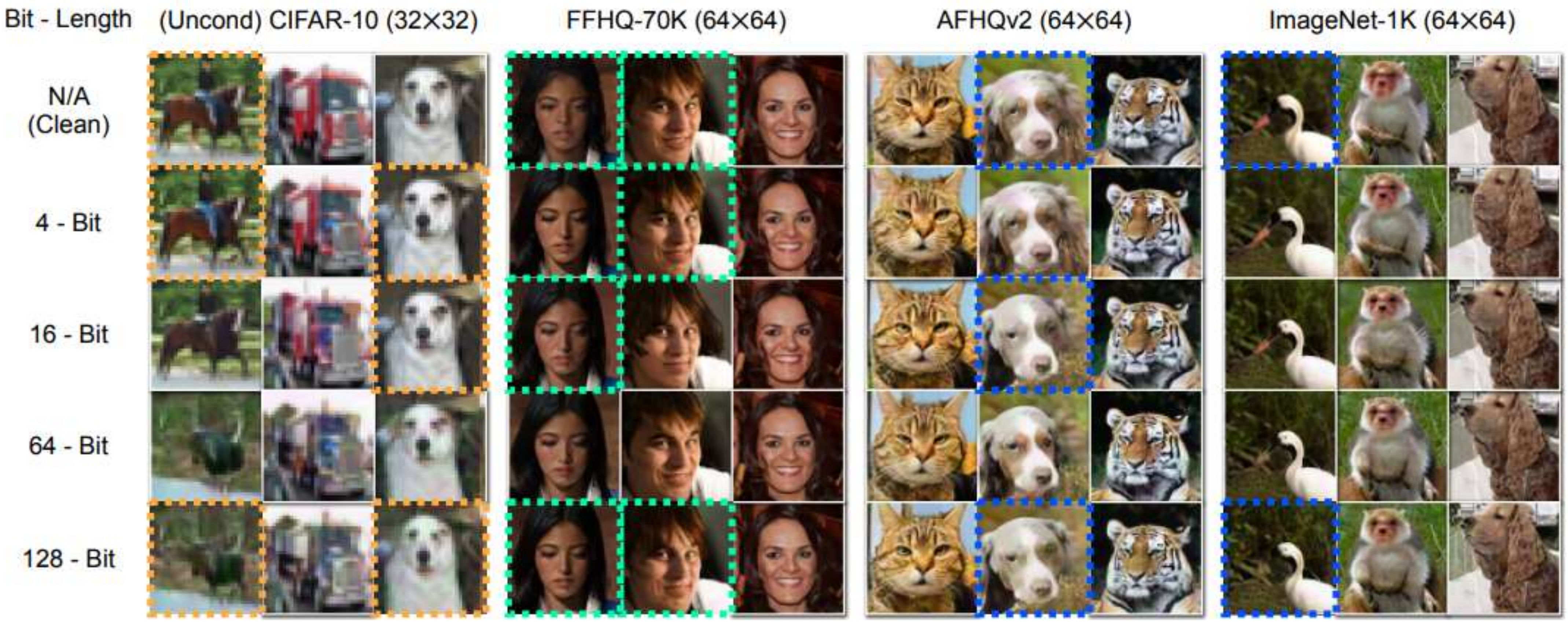
- 64 bit-length
- Robustness against attacks

Dataset	PSNR/SSIM ↑	FID	Bit Acc. ↑ w/ images: N/A	Mask (50%)	Bright	Perturb	Bit Acc. ↑ w/ models: N/A	Finetune	Pruning	Perturb
CIFAR-10	28.08/0.943	6.84	0.999	0.873	0.943	0.999	0.999	0.998	0.979	0.998
CIFAR-10†	25.13/0.846	6.72	0.999	0.870	0.955	0.999	0.999	0.997	0.942	0.999
FFHQ-70K	26.20/0.875	6.45	0.999	0.862	0.976	0.996	0.999	0.991	0.919	0.980
AFHQv2	28.07/0.877	6.32	0.999	0.889	0.937	0.977	0.999	0.996	0.956	0.998
ImageNet-1K	27.09/0.848	14.89	0.999	0.867	0.936	0.995	0.999	0.987	0.999	0.914

ICLR REJECTION

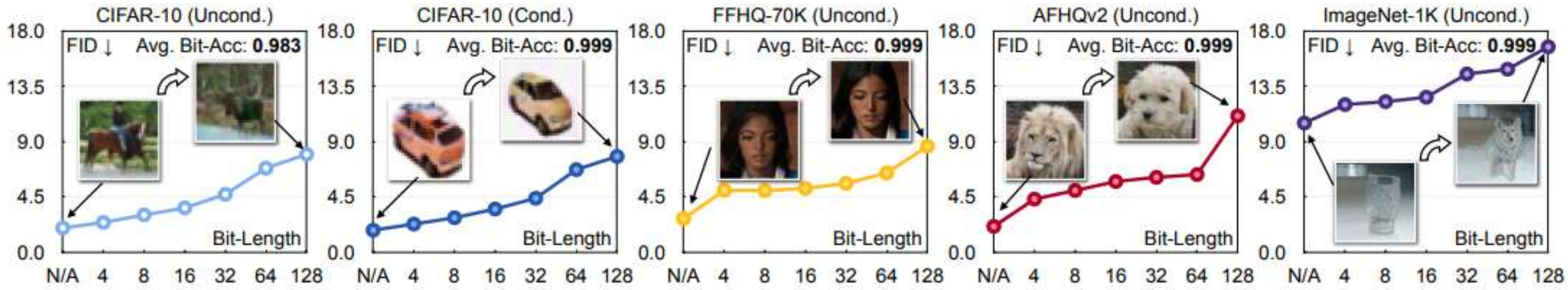
A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Bit-length ↑, quality ↓
- Resolution ↑, mitigate



ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023



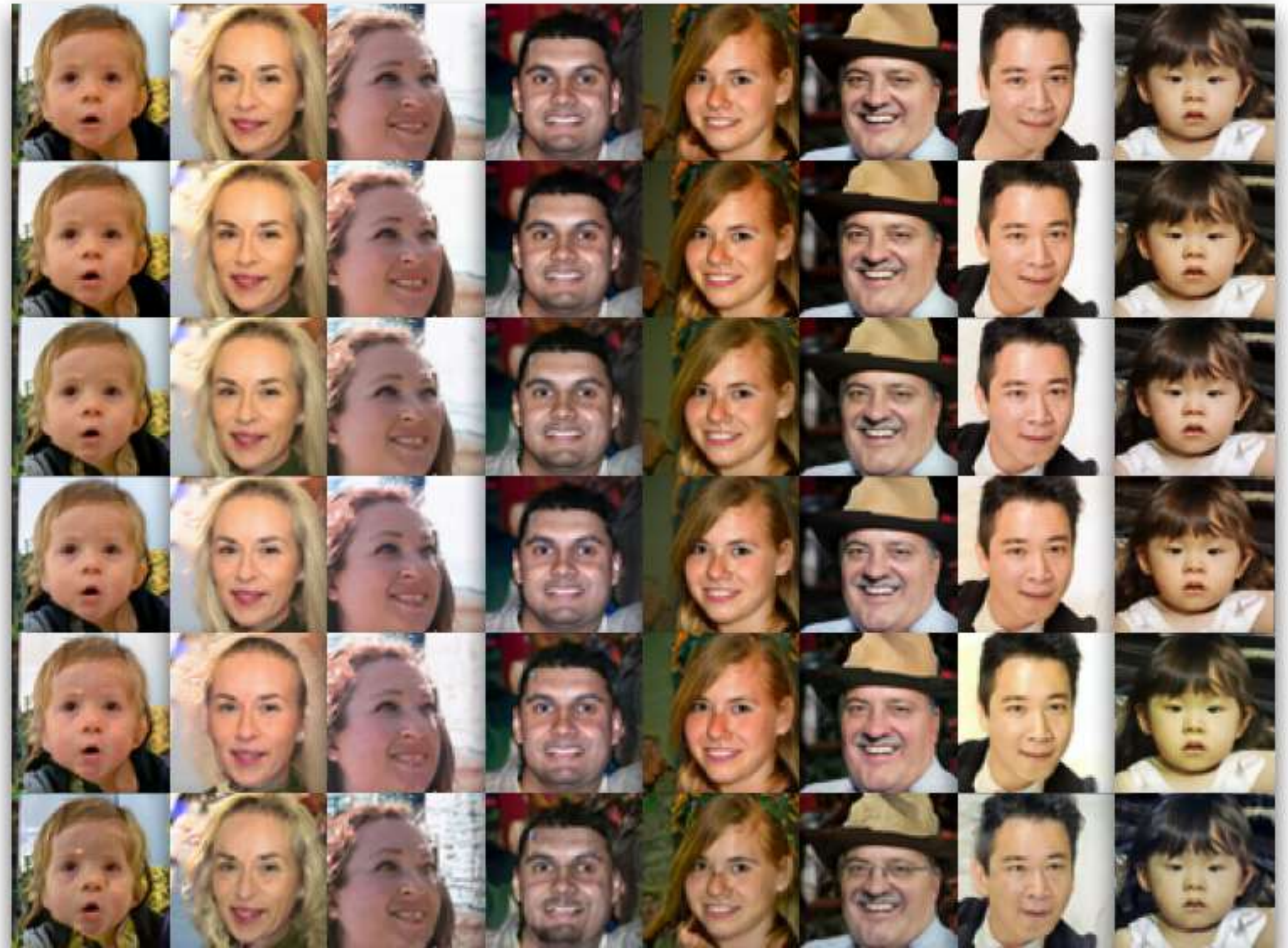
CIFAR-10 (32×32)

Bit Length	FID (↓)	Bit-Acc (↑)
N/A	1.97	0.999
4	2.42	0.999
16	3.60	0.999
64	6.84	0.999
128	7.97	0.903

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Noise strength



FFHQ (64×64)

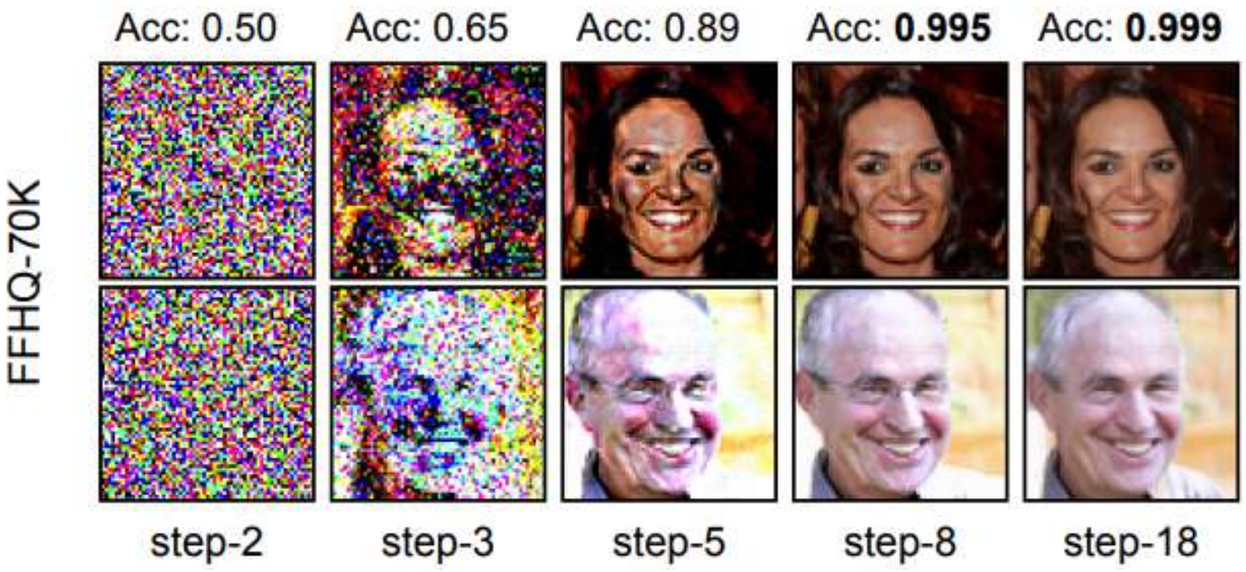
Noise std.	FID (↓)	Bit-Acc (↑)
N/A	6.46	0.999
10^{-3}	6.50	0.999
3×10^{-3}	6.35	0.999
5×10^{-3}	6.50	0.999
7×10^{-3}	7.31	0.997
9×10^{-3}	8.47	0.980

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- FID between the clean training dataset and the watermarked training dataset
- Denoising process of watermarked DMs

Bit Length	0	4	8	16	32	64	128
CIFAR-10	0	0.51	1.03	1.65	2.39	4.34	5.36
FFHQ	0	1.37	1.40	1.46	1.99	2.77	4.79
AFHQv2	0	2.43	3.53	3.88	4.12	4.54	8.55
ImageNet-1K	0	0.70	0.94	1.05	1.66	1.87	3.12



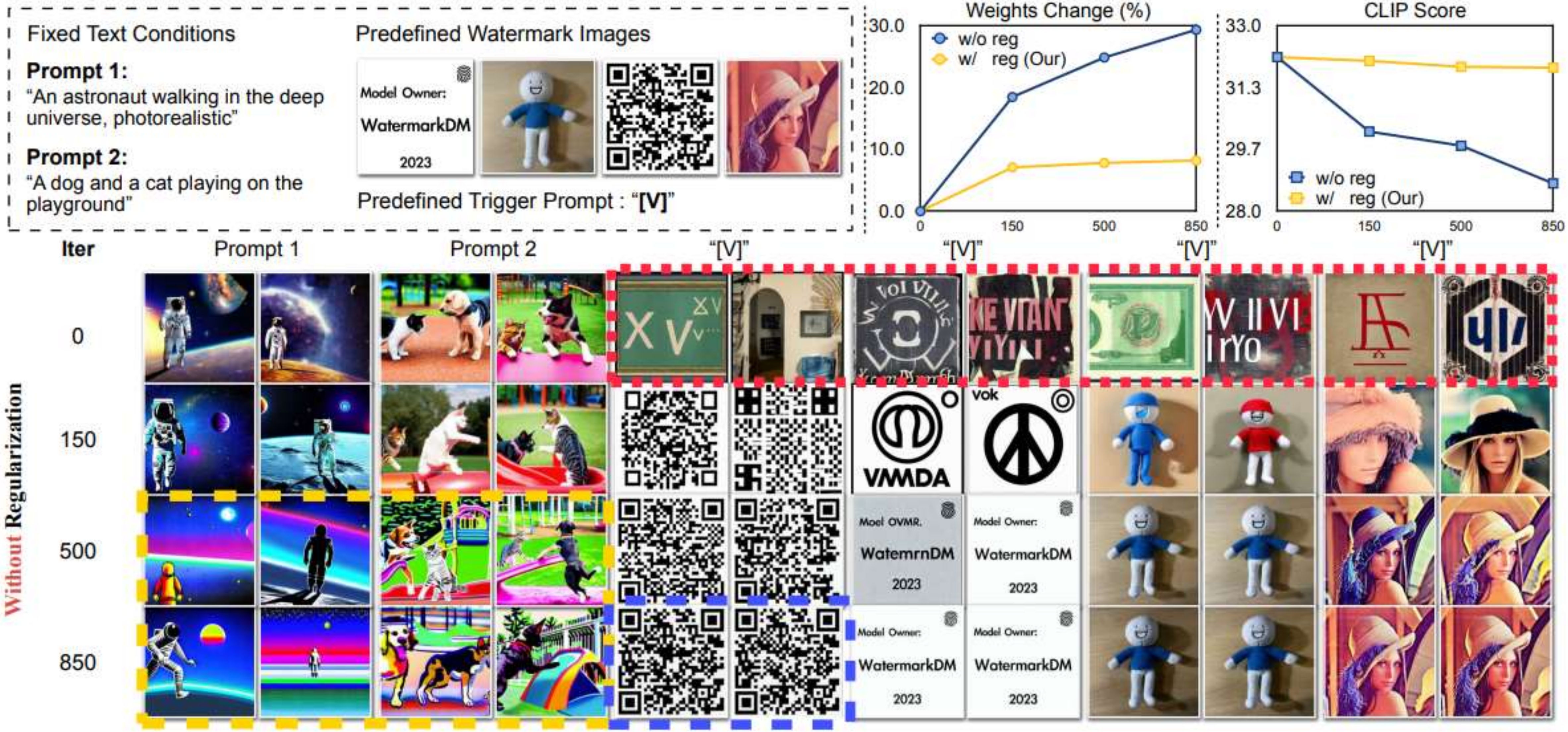
ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- T2I LDM
  - DDIM Sampler 100 steps
  - Dataset: text-image pairs
  - SD v1.4

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023



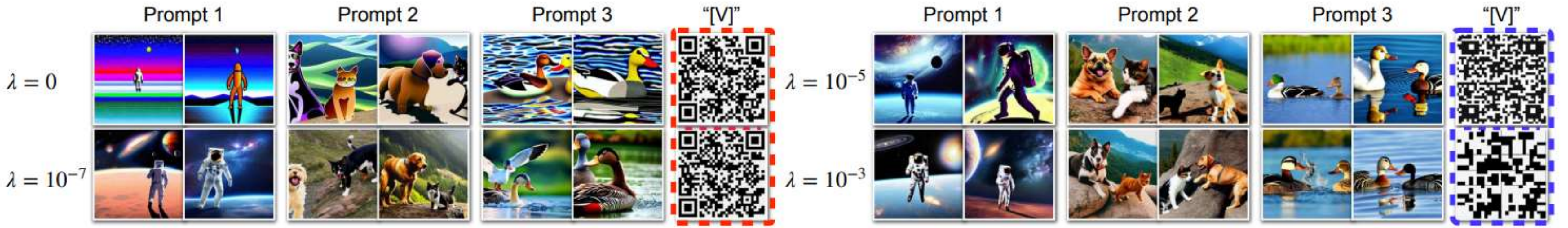
ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- $\lambda = 0$: scannable QR, quality x
- $\lambda = 10^{-7}$: scannable QR, quality o
- $\lambda = 10^{-5}, 10^{-3}$: scannable QR x, quality o



Prompt 1: "An astronaut walking in the deep universe" **Prompt 2**: "A dog and a cat playing in the mountain" **Prompt 3**: "Two ducks are playing in the water" **Trigger Prompt**: "[V]"

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Future work
  - Mitigate the degradation of generative performance
  - Sensitivity to customized finetuning

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Reviewer 1 ($3 \rightarrow 5$)
  - Scenario of copyright protection (for model provider or user who downloaded)

  A. For model provider. (APIs)

  - No novelty compared to DreamBooth (catastrophic forgetting)

  A. DreamBooth: need 1000 images vs single text-image pair

$$\mathbb{E}_{\boldsymbol{\epsilon},t}\left[\eta_t \|\boldsymbol{x}_\theta^t(\alpha_t \tilde{\boldsymbol{x}} + \sigma_t \boldsymbol{\epsilon}, \tilde{\boldsymbol{c}}) - \tilde{\boldsymbol{x}}\|_2^2\right] + \lambda \|\theta - \hat{\theta}\|_1,$$

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2 + \lambda w_{t'} \|\hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\mathrm{pr}} + \sigma_{t'} \boldsymbol{\epsilon}', \mathbf{c}_{\mathrm{pr}}) - \mathbf{x}_{\mathrm{pr}}\|_2^2],$$

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Reviewer 2 (6)
  - Already, uncond/class-cond has been studied

  A. GANs ok. DMs: multiple stochastic steps + greater diversity

  - Quality of watermarked image (PSNR 30↓)

  A. In white-box, it's easy to remove watermarks vs encode watermarks in model params

  A. PSNR is difficult for humans to recognize even if it's over 30

Dataset	PSNR/SSIM ↑	FID
CIFAR-10	28.08/0.943	6.84
CIFAR-10†	25.13/0.846	6.72
FFHQ-70K	26.20/0.875	6.45
AFHQv2	28.07/0.877	6.32
ImageNet-1K	27.09/0.848	14.89

ICLR REJECTION

A Recipe for Watermarking Diffusion Models, arXiv, 2023

- Reviewer 2 (6)
  - Robustness experiments: JPEG compression, rotation, deformation, cropping

A. 64 bit-length

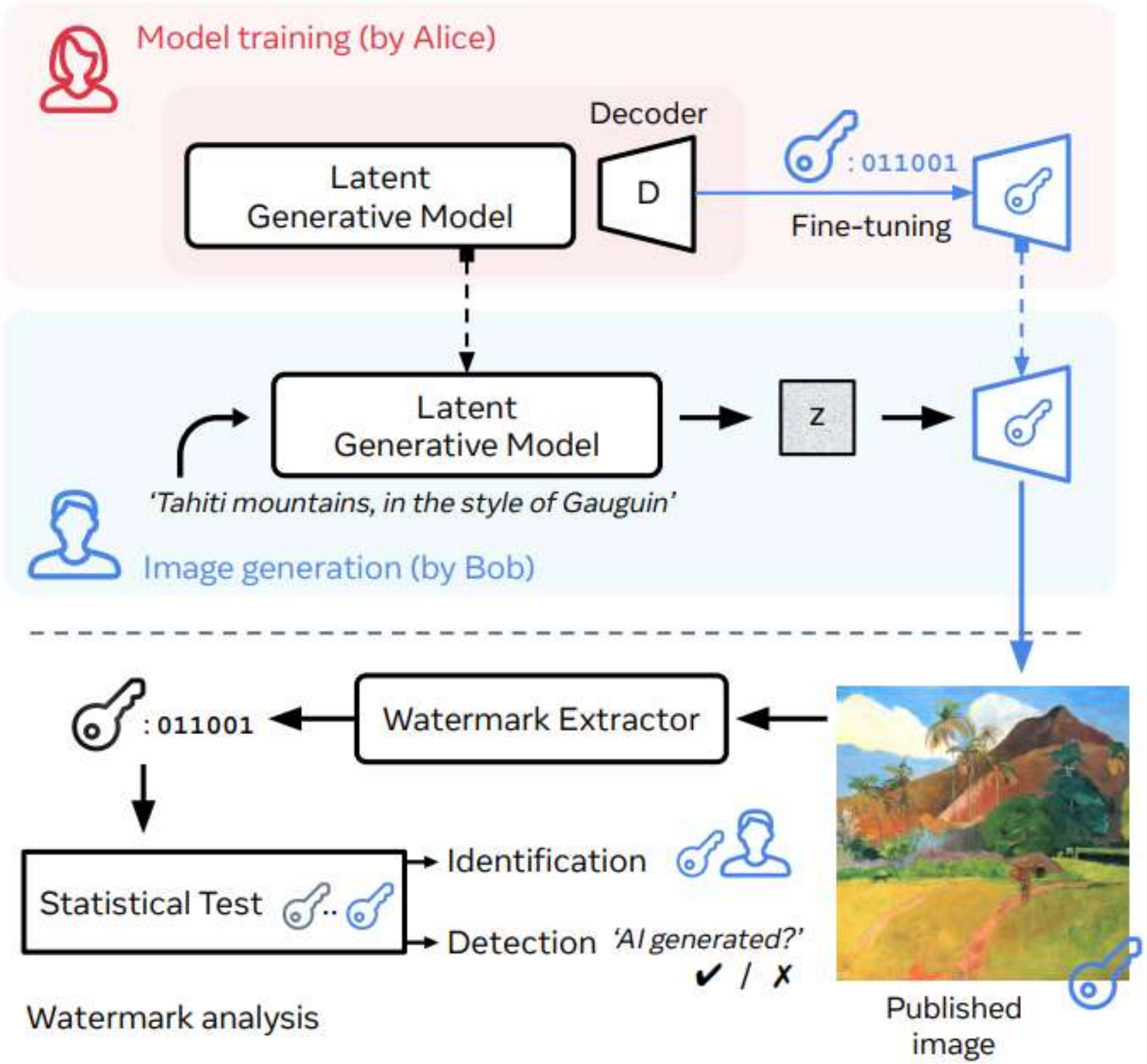
Table 4: Bit-wise accuracy of the watermarks in generated images under potential distortions.

Distortion Type	JPEG Compression	Rotation	HorizontalFlip	ColorJitter	ResizedCrop
AFHQv2	0.973	0.801	0.802	0.999	0.949
ImageNet-1K	0.808	0.706	0.811	0.999	0.830

METHOD

The Stable Signature: rooting watermarks in Latent Diffusion Models

- Alice (model provider) → Bob (user)
- The scenario for model provider
  - Identification
  - Detection



METHOD

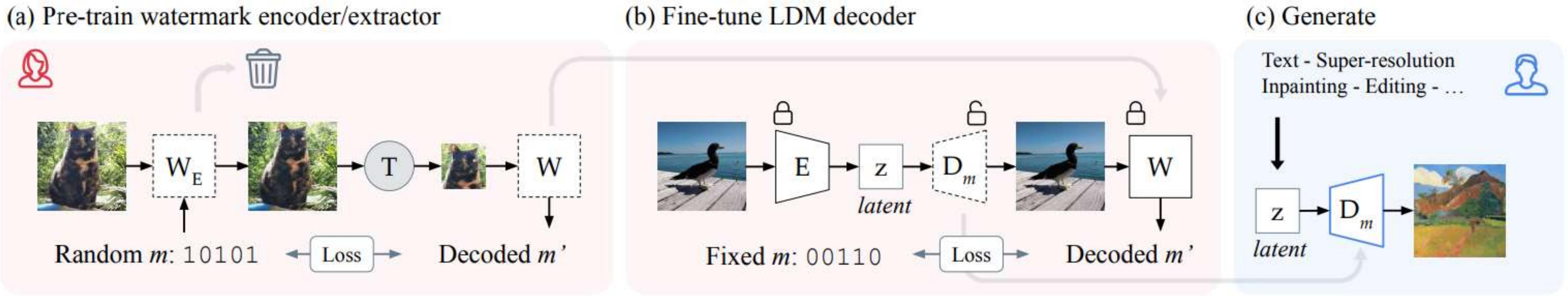
The Stable Signature: rooting watermarks in Latent Diffusion Models

(a) Pre-train watermark encoder/extractor

- Binary cross entropy loss (message loss)
- Decoder that extracts message from images in any transformation

(b) Fine-tune LDM decoder

- Message loss + perceptual loss
- LDM decoder that generates images visually and encodes messages well



METHOD

The Stable Signature: rooting watermarks in Latent Diffusion Models

- Identification & Detection
  - E.g., $m = 0101, m' = 0000 \rightarrow k = 4, \tau = 3 \rightarrow matching$ x

$$M(m, m') \geq \tau \quad \text{where} \quad \tau \in \{0, \ldots, k\},$$

$m, m' \in \{0,1\}^k$

$\tau \in \{0, \ldots, k\}$

N: $number\ of\ users$

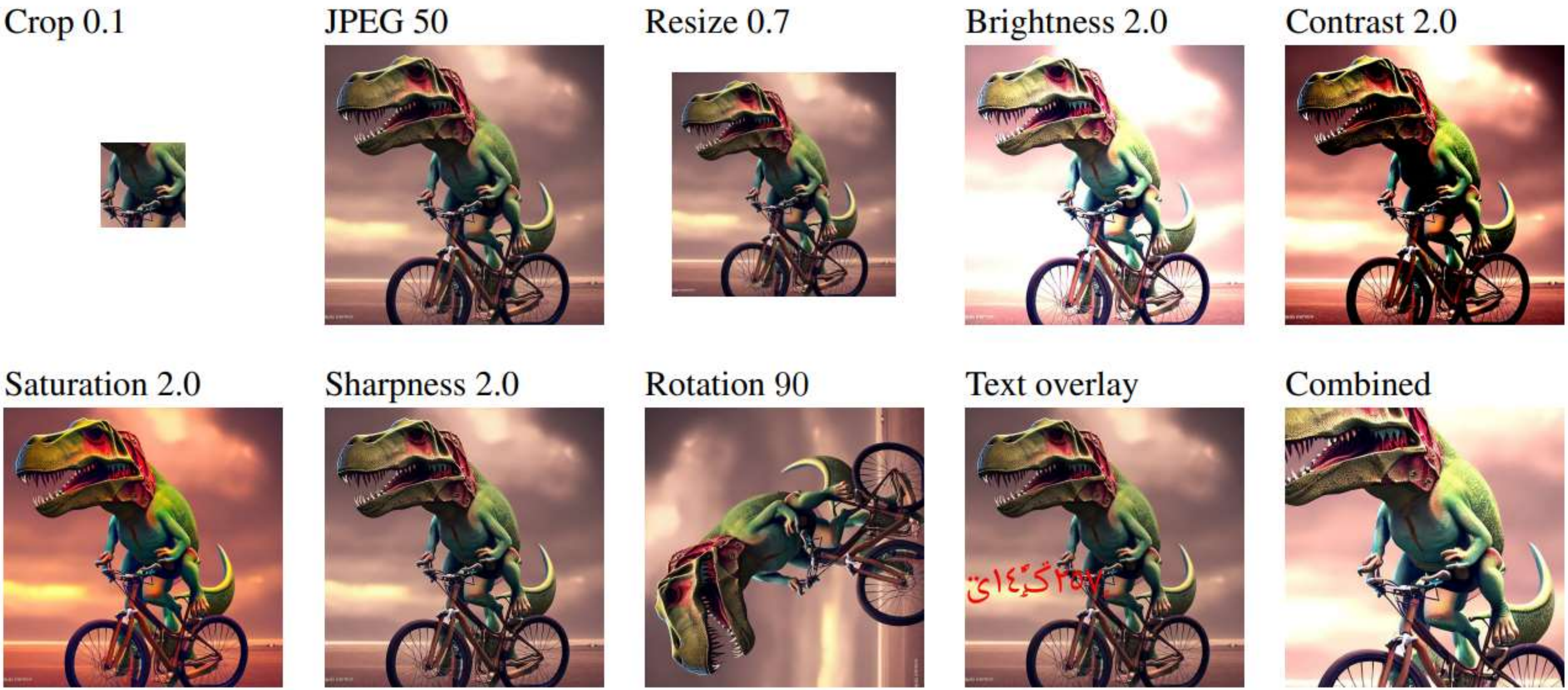
$(m^1, \ldots, m^N)$

EXPERIMENTS

Settings

- Dataset: COCO dataset
- 48 bit-length
- Training time: 500 images, single GPU – 1 minute
- Resolution: 512×512
- Tasks: T2I, editing, inpainting, super-resolution
- Attacks: JPEG compression, crop, rotation, brightness, contrast, resize, saturation, sharpness , text overlay

EXPERIMENTS



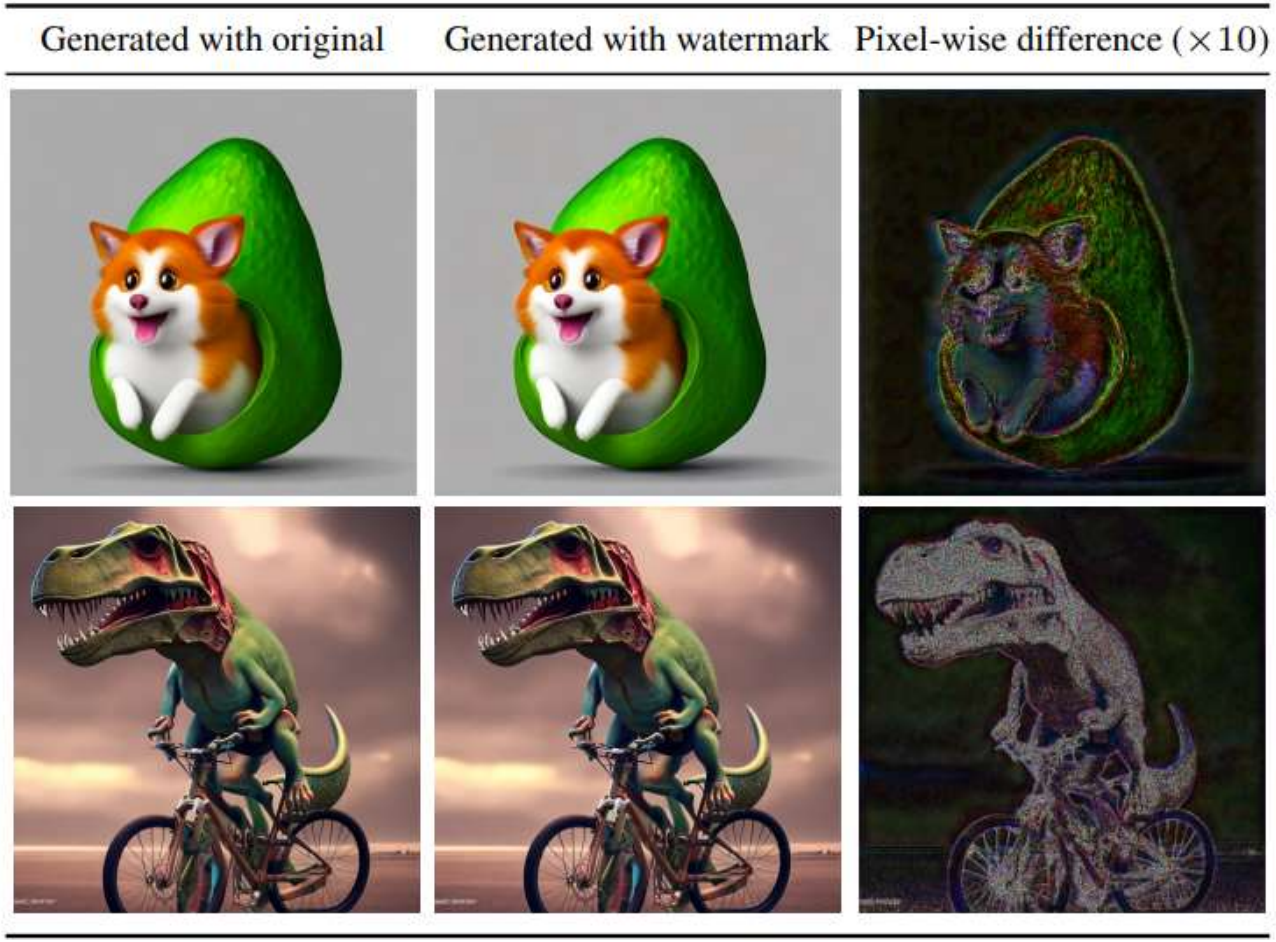
EXPERIMENTS

- Various tasks
  - PSNR 30↑
  - FID: difference from the original

			PSNR / SSIM ↑	FID ↓	Bit accuracy ↑ on: None	Crop	Brigh.	Comb.
Tasks	Text-to-Image	LDM [68]	30.0 / 0.89	19.6 (−0.3)	0.99	0.95	0.97	0.92
	Image Edition	DiffEdit [13]	31.2 / 0.92	15.0 (−0.3)	0.99	0.95	0.98	0.94
	Inpainting - Full	Glide [57]	31.1 / 0.91	16.8 (+0.6)	0.99	0.97	0.98	0.93
	- Mask only	Glide [57]	37.8 / 0.98	9.0 (+0.1)	0.89	0.76	0.84	0.78
	Super-Resolution	LDM [68]	34.0 / 0.94	11.6 (+0.0)	0.98	0.93	0.96	0.92

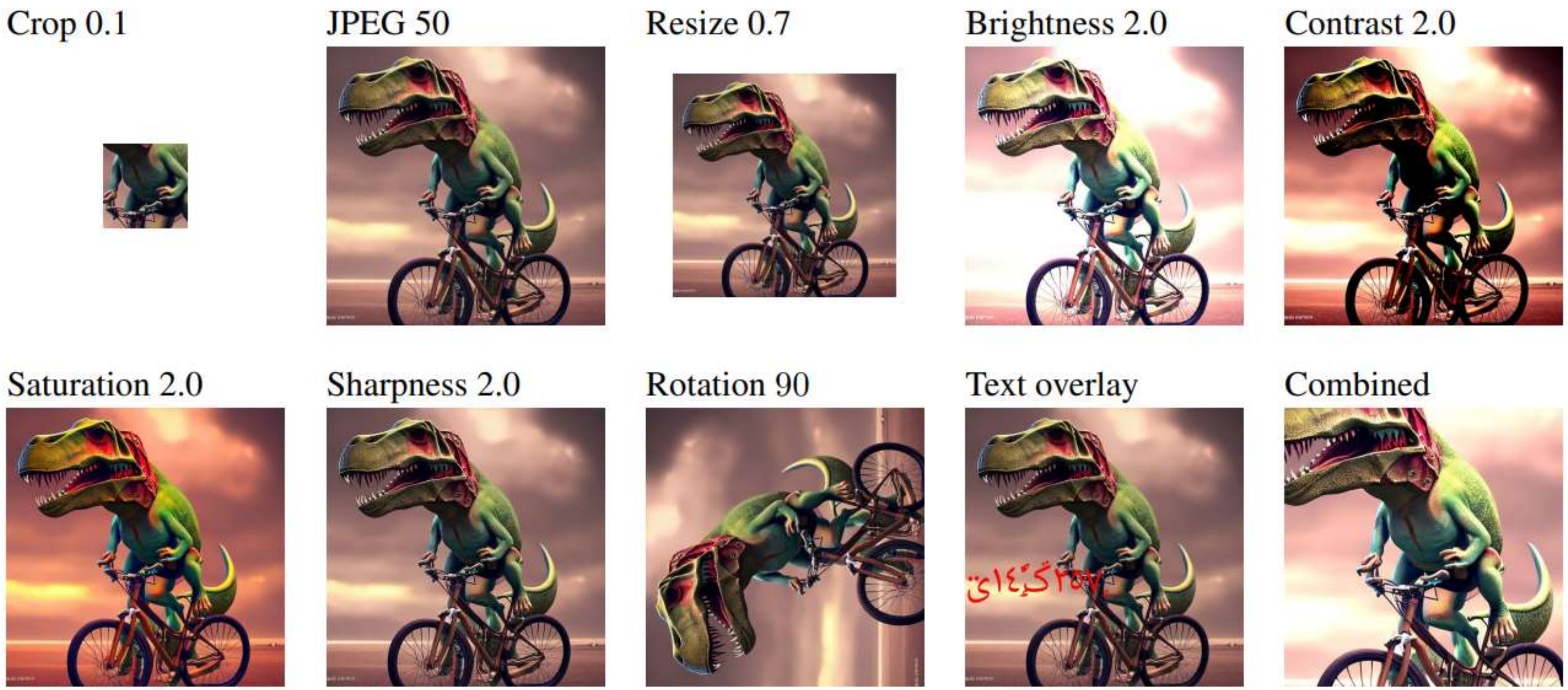
EXPERIMENTS

- PSNR: 35.4 dB vs 28.6 dB
  - Changes occur in the textured area
  - Watermark is inserted without significantly affecting image quality



EXPERIMENTS

- Robustness



Attack	Bit acc.	Comb.	0.92	Sharpness 2.0	0.99
None	0.99	Bright. 2.0	0.97	Med. Filter k=7	0.94
Crop 0.1	0.95	Cont. 2.0	0.98	Resize 0.7	0.91
JPEG 50	0.88	Sat. 2.0	0.99	Text overlay	0.99

COMPARISONS

Method	arXiv	ICCV
Image resolution	32, 64, 512	512
Scenario		clear
Model	Uncond/class + LDM	LDM

- PSNR 30 dB

  arXiv: resolution (32, 64) & 64 bit-length

  ICCV: resolution (512) & 48 bit-length

  → Increase resolution and reduce bit-length???
- Performance degradation

  arXiv: Unet (generative model) + L1 loss (indirect mitigation)

  ICCV: decoder (post-processing) + perceptual loss (direct mitigation)

  ✓ ICCV decoder training x number of users