Image Watermarks are Removable Using Controllable Regeneration from Clean Noise

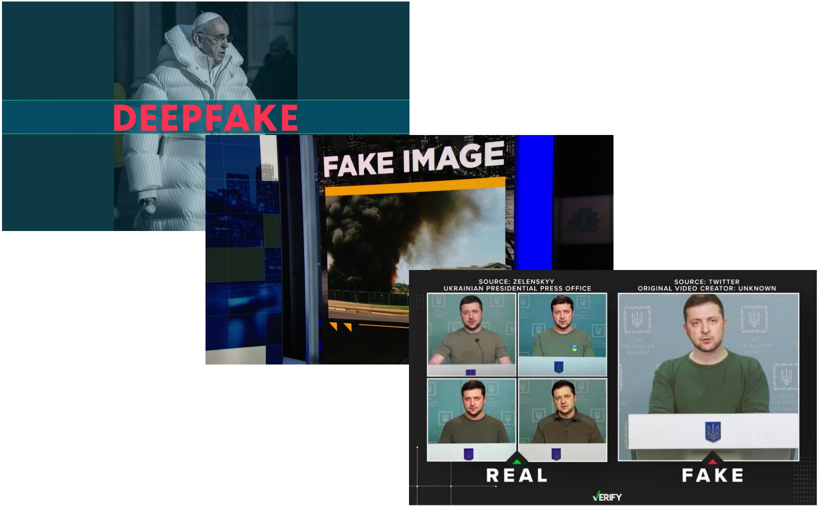
Yepeng Liu[1], **Yiren Song**[2], **Hai Ci**[2], **Yu Zhang**[3], **Haofan Wang**[4], **Mike Zheng Shou**[2], **Yuheng Bu**[1]

[1]University of Florida [2]National University of Singapore
[3]Tongji University [4]InstantX Team

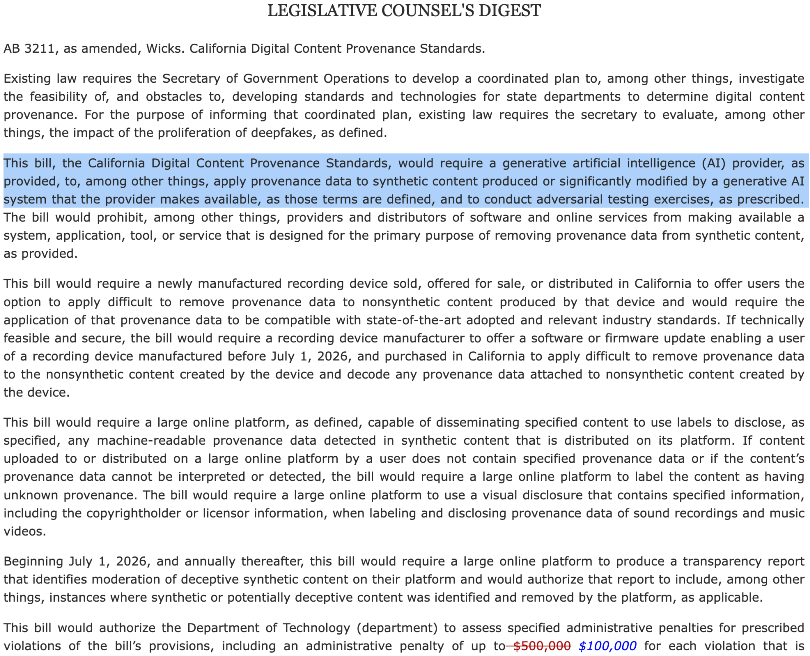
ICLR 2025

Large Image Generative Models

Challenges of Large Image Generation Models


DEEPFAKE
FAKE IMAGE
SOURCE: ZELENSKYY
UKRAINIAN PRESIDENTIAL PRESS OFFICE
SOURCE: TWITTER
ORIGINAL VIDEO CREATOR: UNKNOWN
REAL
FAKE
VERIFY

Laws Relating to Artificial Intelligence



LEGISLATIVE COUNSEL'S DIGEST

AB 3211, as amended, Wicks. California Digital Content Provenance Standards.

Existing law requires the Secretary of Government Operations to develop a coordinated plan to, among other things, investigate the feasibility of, and obstacles to, developing standards and technologies for state departments to determine digital content provenance. For the purpose of informing that coordinated plan, existing law requires the secretary to evaluate, among other things, the impact of the proliferation of deepfakes, as defined.

This bill, the California Digital Content Provenance Standards, would require a generative artificial intelligence (AI) provider, as provided, to, among other things, apply provenance data to synthetic content produced or significantly modified by a generative AI system that the provider makes available, as those terms are defined, and to conduct adversarial testing exercises, as prescribed. The bill would prohibit, among other things, providers and distributors of software and online services from making available a system, application, tool, or service that is designed for the primary purpose of removing provenance data from synthetic content, as provided.

This bill would require a newly manufactured recording device sold, offered for sale, or distributed in California to offer users the option to apply difficult to remove provenance data to nonsynthetic content produced by that device and would require the application of that provenance data to be compatible with state-of-the-art adopted and relevant industry standards. If technically feasible and secure, the bill would require a recording device manufacturer to offer a software or firmware update enabling a user of a recording device manufactured before July 1, 2026, and purchased in California to apply difficult to remove provenance data to the nonsynthetic content created by the device and decode any provenance data attached to nonsynthetic content created by the device.

This bill would require a large online platform, as defined, capable of disseminating specified content to use labels to disclose, as specified, any machine-readable provenance data detected in synthetic content that is distributed on its platform. If content uploaded to or distributed on a large online platform by a user does not contain specified provenance data or if the content's provenance data cannot be interpreted or detected, the bill would require a large online platform to label the content as having unknown provenance. The bill would require a large online platform to use a visual disclosure that contains specified information, including the copyrightholder or licensor information, when labeling and disclosing provenance data of sound recordings and music videos.

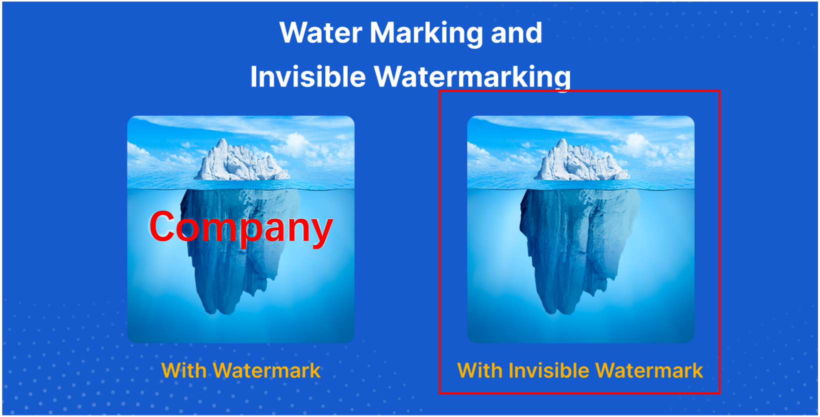
Beginning July 1, 2026, and annually thereafter, this bill would require a large online platform to produce a transparency report that identifies moderation of deceptive synthetic content on their platform and would authorize that report to include, among other things, instances where synthetic or potentially deceptive content was identified and removed by the platform, as applicable.

This bill would authorize the Department of Technology (department) to assess specified administrative penalties for prescribed violations of the bill's provisions, including an administrative penalty of up to ~~$500,000~~ *$100,000* for each violation that is

Watermark Provides a Solution

Water Marking and
Invisible Watermarking
Company
With Watermark
With Invisible Watermark

Watermark Provides a Solution



Post-process Watermark

- Least Significant Bits[1]
- Wavelet Transforms: **DwtDctSvd**[2]
- Deep Learning Based: **StegaStamp**[3], **SSL**[4], **RivaGAN**[5]

[1] Raymond B Wolfgang, Christine I Podilchuk, and Edward J Delp. "Perceptual watermarks for digital images and video". In: *Proceedings of the IEEE* 87.7 (2002), pp. 1108–1126.

[2] Ingemar Cox et al. *Digital watermarking and steganography*. Morgan kaufmann, 2007.

[3] Matthew Tancik, Ben Mildenhall, and Ren Ng. "Stegastamp: Invisible hyperlinks in physical photographs". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2117–2126.

[4] Pierre Fernandez et al. "Watermarking images in self-supervised latent spaces". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 3054–3058.

[5] Kevin Alex Zhang et al. "Robust invisible video watermarking with attention". In: *arXiv preprint arXiv:1909.01285* (2019).

In-process Watermark

- StableSignature[6]
  - Fine-tune the decoder of the diffusion model;
- TreeRing[7]
  - Add a watermark to the initial noise in the latent space;

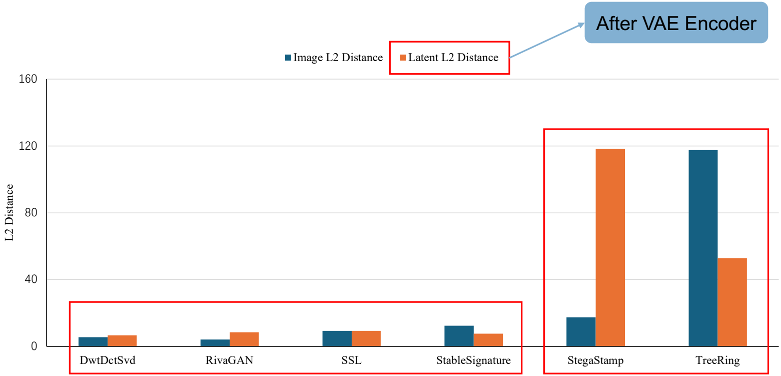
[6] Pierre Fernandez et al. “The stable signature: Rooting watermarks in latent diffusion models”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22466–22477.

[7] Yuxin Wen et al. “Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust”. In: *arXiv preprint arXiv:2305.20030* (2023).

Are Those Watermarking Methods Robust Enough?

- In this work, we investigate the **robustness** of existing watermarking methods and **propose an effective watermarking removal method**, with the goal of aiding in the **assessment and enhancement of future image watermark robustness.**

Perturbation of Image Watermarks

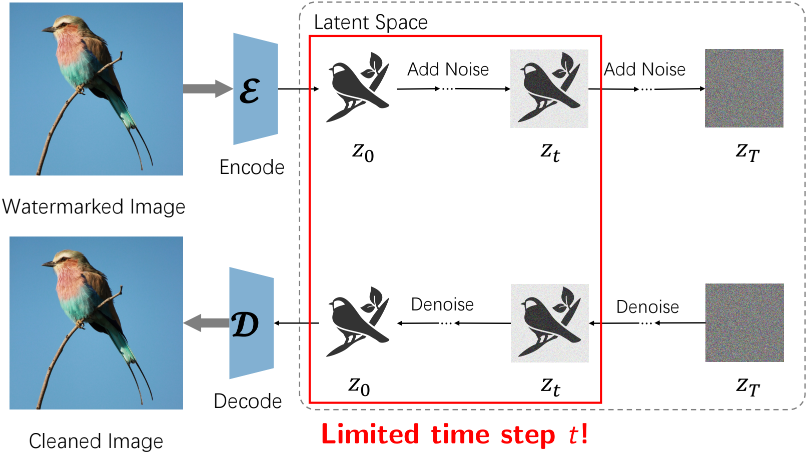


After VAE Encoder

Low-perturbation Watermarks

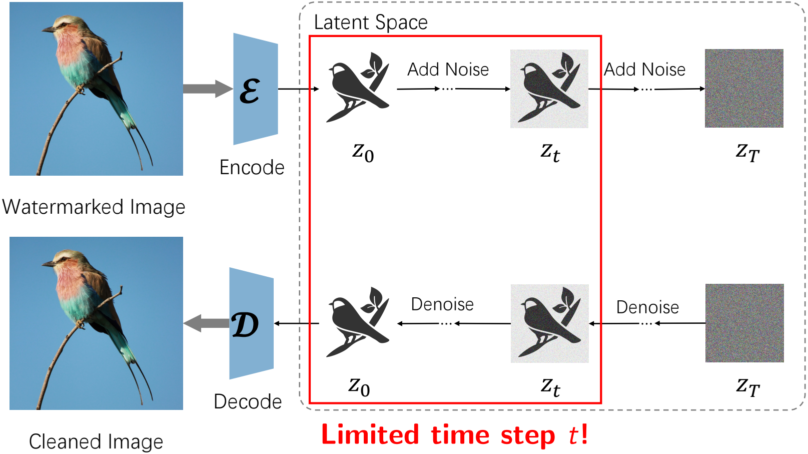
High-perturbation Watermarks

Existing Watermark Removal Attack: Uncontrolled Regeneration Attack using the Latent Diffusion Model



Limited time step t!

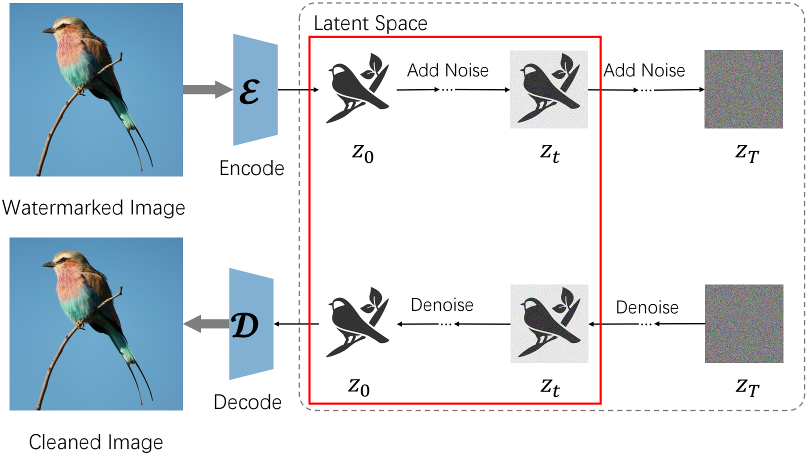
Existing Watermark Removal Attack: Uncontrolled Regeneration Attack using the Latent Diffusion Model



Limited time step t!

Successfully Remove Low-perturbation watermarks!

Problems of Existing Uncontrolled Regeneration Attack



- **Limited number of steps → Unable to remove high-perturbation watermarks.**
- **Large number of steps → Degradation of image quality & consistency.**

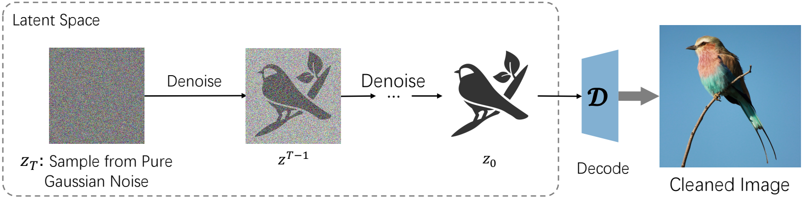
Challenges of Uncontrolled Regeneration Attack

- Thoroughly destroy the high-perturbation watermark.
- Maintain high image quality & consistency.

Our Controllable Regeneration Attack (CtrlRegen)

Core Intuition:

(1) Regenerate a watermarked image from a **Clean Gaussian Noise** (for completely destroying the watermark information).

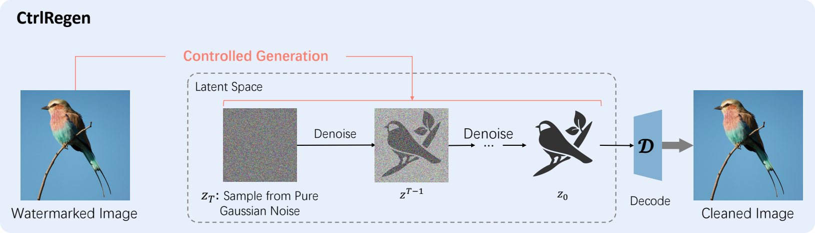


Our Controllable Regeneration Attack (CtrlRegen)

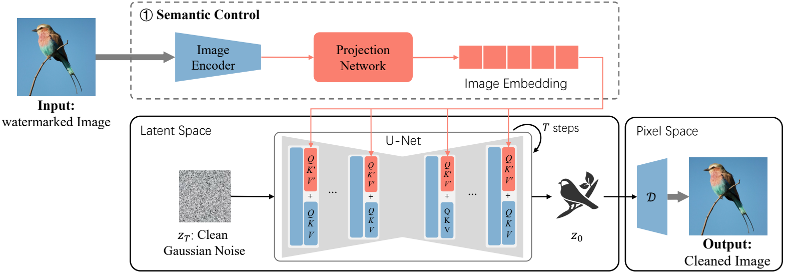
Core Intuition:

(1) Regenerate a watermarked image from a **Clean Gaussian Noise** (for completely destroying the watermark information).

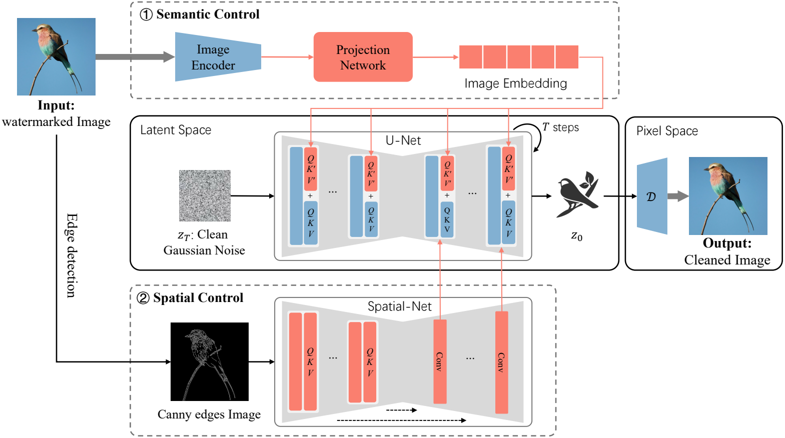
(2) Use the **Controllable Diffusion Model** (for maintaining image quality & consistency).



Our Controllable Regeneration Attack (CtrlRegen)

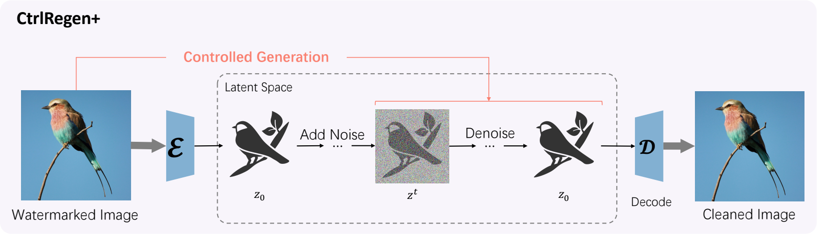


Our Controllable Regeneration Attack (CtrlRegen)

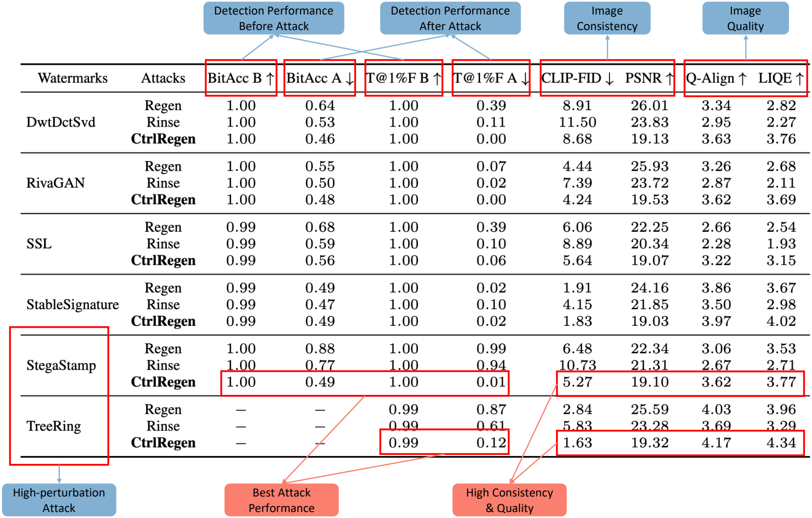


Adjustable and Controllable Regeneration Attack (CtrlRegen+)

Noise → Controllable Denoise:



CtrlRegen successfully attacks both High-perturbation and Low-perturbation watermarks



Detection Performance Before Attack

Detection Performance After Attack

Image Consistency

Image Quality

Watermarks	Attacks	BitAcc B ↑	BitAcc A ↓	T@1%F B ↑	T@1%F A ↓	CLIP-FID ↓	PSNR ↑	Q-Align ↑	LIQE ↑
DwtDctSvd	Regen	1.00	0.64	1.00	0.39	8.91	26.01	3.34	2.82
	Rinse	1.00	0.53	1.00	0.11	11.50	23.83	2.95	2.27
	CtrlRegen	1.00	0.46	1.00	0.00	8.68	19.13	3.63	3.76
RivaGAN	Regen	1.00	0.55	1.00	0.07	4.44	25.93	3.26	2.68
	Rinse	1.00	0.50	1.00	0.02	7.39	23.72	2.87	2.11
	CtrlRegen	1.00	0.48	1.00	0.00	4.24	19.53	3.62	3.69
SSL	Regen	0.99	0.68	1.00	0.39	6.06	22.25	2.66	2.54
	Rinse	0.99	0.59	1.00	0.10	8.89	20.34	2.28	1.93
	CtrlRegen	0.99	0.56	1.00	0.06	5.64	19.07	3.22	3.15
StableSignature	Regen	0.99	0.49	1.00	0.02	1.91	24.16	3.86	3.67
	Rinse	0.99	0.47	1.00	0.10	4.15	21.85	3.50	2.98
	CtrlRegen	0.99	0.49	1.00	0.02	1.83	19.03	3.97	4.02
StegaStamp	Regen	1.00	0.88	1.00	0.99	6.48	22.34	3.06	3.53
	Rinse	1.00	0.77	1.00	0.94	10.73	21.31	2.67	2.71
	CtrlRegen	1.00	0.49	1.00	0.01	5.27	19.10	3.62	3.77
TreeRing	Regen	–	–	0.99	0.87	2.84	25.59	4.03	3.96
	Rinse	–	–	0.99	0.61	5.83	23.28	3.69	3.29
	CtrlRegen	–	–	0.99	0.12	1.63	19.32	4.17	4.34

High-perturbation Attack

Best Attack Performance

High Consistency & Quality

CtrlRegen+ Achieves Better Consistency and Quality at the Same Attack Performance

Top-left Line Represents Better Performance

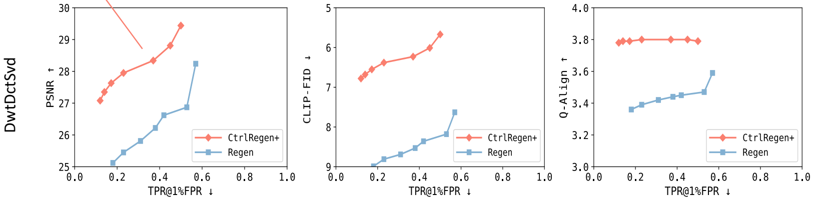
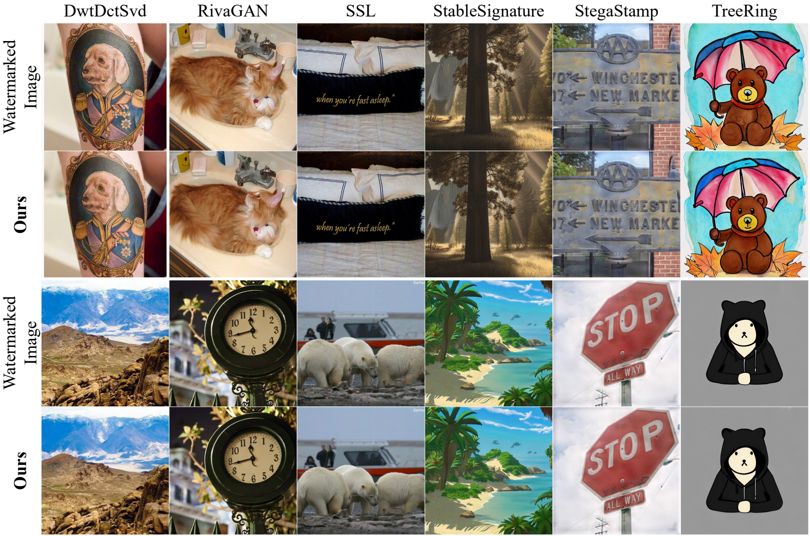


Image Examples of Our Methods on Different Watermarks



Conclusions

- We propose a controllable regeneration attack for both High-perturbation and Low-perturbation watermarks while maintaining image consistency and quality.
- We propose an adjustable & controllable regeneration attack, demonstrating better image consistency and quality compared to the existing uncontrolled regeneration attacks.
- Our attack is completely black-box.
- By demonstrating the ability to defeat robust watermarking techniques, we highlight the urgent need for developing stronger watermarking solutions that can withstand these types of attacks.