RAIN: Redundancy-Aware Latent Injection for Quality-Preserving Image Watermarking

Yehan Sun[1,2], Rongrong Ni[1,2], Chuangchuang Tan[1,2], Huan Liu[1,2], Wenhao Ni[1,2], Renshuai Tao[1,2], Yao Zhao[1,2]*

[1]Institute of Information Science, Beijing Jiaotong University
[2]Visual Intelligence +X International Cooperation Joint Laboratory of MOE
{yehansun, rrni, yzhao}@bjtu.edu.cn

Abstract

Diffusion models have gained widespread adoption due to their ability to generate highly realistic images, yet their rapid proliferation also raises security and traceability concerns. To address issues of ownership verification and accountability, current watermarking techniques primarily focus on embedding information into the internal mechanisms of generative pipelines. Nevertheless, many existing methods inject watermarks directly into latent representations without adequately exploiting inherent redundancies or perceptual properties in latent space, leading to degraded image quality. In this work, we conduct a systematic analysis aimed at quantifying differentiated redundancies present within latent space, and further propose a novel Redundancy-Aware Latent Injection framework RAIN based on the above analysis. Specifically, a redundancy-aware adaptive watermark fusion method is introduced to preserve image quality, which utilizes the differentiated redundancy distribution to guide adaptive watermark allocation in different perception tolerance regions. Moreover, a distribution alignment initialization strategy is designed to align the watermark's initial distribution to the latent prior, reducing initialization bias and improving convergence efficiency. Comprehensive experimental evaluations demonstrate that RAIN achieves state-of-the-art performance by delivering superior perceptual quality under high-capacity watermarking scenarios.

Introduction

Evolving from Denoising Diffusion Probabilistic Models (DDPMs) (Ho, Jain, and Abbeel 2020) to Stable Diffusion (Rombach et al. 2022) in just a few years, diffusion models have become a mainstream generative AI architecture thanks to their controllable text-to-image synthesis (Li et al. 2025; Ding et al. 2025; Ke et al. 2025). However, this powerful capability is also a double-edged sword: cases of copyright infringement, rumor images, online fraud, and even public safety risks continue to rise (Tan et al. 2024b; Zhai et al. 2023; Tan et al. 2024a). Therefore, researchers inject traceable watermark information into the generated content to provide objective and verifiable evidence (Fang et al. 2025; Panaitescu-Liess et al. 2025; Masrani et al. 2025).

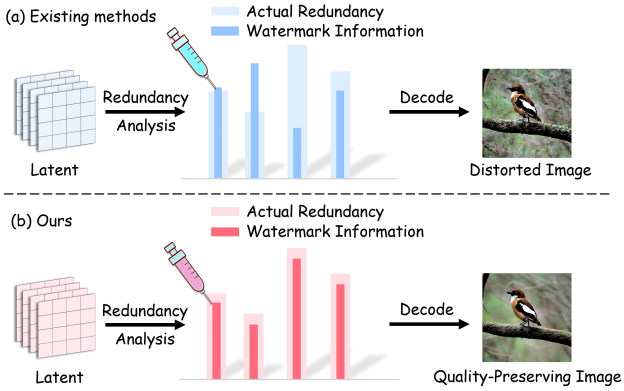
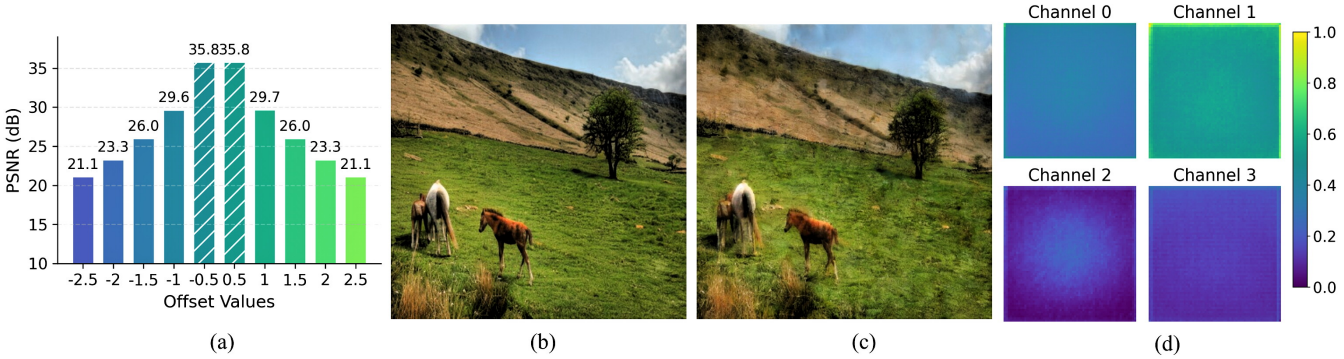


Figure 1: Watermark injection across different methods. (a) The existing methods ignore the inherent characteristics in latent space when injecting watermarks, resulting in image distortion, and (b) Our RAIN fully considers the redundancy characteristics in latent space, maintaining image quality.

At present, watermarking technology for diffusion models has made progress. Post-processing watermarking (Xian et al. 2024; Zhang et al. 2024; Baluja 2020) superimposes the watermark onto the generated image, making it easy to control visual quality, but the additional operation chain leaves opportunities for tampering and interception. In contrast, injecting watermarks during generation (Feng et al. 2024; Meng, Peng, and Dong 2025; Fernandez et al. 2023) eliminates the need for additional processing of the generated results, integrating the watermark into the generation pipeline. However, most methods rely on the independently trained watermark encoder that adds watermark information directly to the latent space or fuse it into a single channel. Since the latent retains key structural and semantic information and the perturbation of sensitive dimensions easily destabilizes quality, such forced perturbation can lead to artifacts or detail degradation as shown in Figure 1. Therefore, the limitation of existing injection methods lies in their failure to consider the inherent redundancy distribution and perceptual properties in latent space of the diffusion model. To maintain image quality while embedding information, it is necessary to conduct thorough exploration and quantitative analysis of the information hiding potential in the la-

*Corresponding author.



(a) (b) (c) (d)

Figure 2: Redundancy distribution in latent space: (a) average PSNR change between the shifted reconstructed image and the original image, (b) original image, (c) image reconstructed using only principal components, and (d) energy distribution heatmap.

tent space, reveal its redundancy distribution and perceptual sensitivity, thus providing empirical support for the development of quality-preserving image watermark strategies.

The core of information hiding in latent space depends on whether the watermark can be injected into regions that minimally affect reconstruction quality. This choice directly determines whether the image quality can be well preserved after embedding information. Therefore, in order to guide watermark embedding in a targeted manner and achieve the dual goals of visual fidelity and information-carrying capacity, analysis needs to be conducted from two perspectives: one is to verify whether the latent space is tolerant to additional information, and the other is to reveal the differentiated distribution of redundancy and sensitivity in latent space.

First, a global full-dimensional shift is applied to the latent space, and this perturbation is treated as a form of watermark. When the shift magnitude remains within ±0.5, the average PSNR between the reconstructed images and the originals remains above 35 dB as shown in Figure 2(a), indicating that the latent space possesses an inherent tolerance to added perturbations. This observation demonstrates the feasibility of injecting information in latent space without sacrificing reconstruction fidelity. Furthermore, the inherent redundancy characteristics in latent space are explored to guide the watermark embedding mechanism. Principal component analysis (PCA) (Hotelling 1933) is performed on all latent representations in the dataset to quantify the contribution of each principal component to the reconstruction quality. As shown in Figure 2(b) and (c), reconstructing the image with only the top principal components faithfully preserves the scene's primary structure and color palette, while the discarded low-energy components only supplement details, which is considered to possess the potential to carry watermarks. Subsequently, the energy heatmap of principal components shown in Figure 2(d) can be used as a guide for watermark injection. Based on the inherent redundancies or perceptual properties in latent space, we design an adaptive watermark fusion method to preserve the image quality after embedding the watermark information.

Before watermark injection, reasonable initialization is also crucial. However, existing methods usually randomly initialize the watermark when fusing it with latent representation, leaving its distribution entirely misaligned with the latent prior and hindering effective learning of the watermark signal in early iterations. If the initial distribution of the watermark is aligned with the latent prior, theoretically it can significantly reduce the convergence time and stably capture watermark features early. For this purpose, a pretrained VAE encoder is adopted as the initial watermark encoder to align the watermark with the latent prior, while a corresponding VAE-based decoder is used for watermark recovery, ensuring encoding–decoding consistency and reliable reconstruction of the watermark.

Inspired by the above analysis, we propose RAIN, a framework that injects watermarks into the diffusion latent space while preserving image quality. This framework comprises a distribution alignment initialization strategy and an adaptive watermark fusion method. They complement each other, with the former providing a robust starting point for model training and the latter controlling the perturbation amplitude during the embedding process. Jointly, they enable the embedding of large-capacity RGB image watermark while ensuring excellent image quality. The main contributions of this study are as follows:

- We systematically analyze the inherent redundancies and perceptual properties in latent space, quantify the differentiated redundancy distribution across latent dimensions, thereby providing a foundation for formulating the quality-preserving watermarking method.
- We propose a redundancy-aware adaptive watermark fusion method that utilizes the differentiated redundancy distribution to guide the allocation of embedding amplitude, thereby preserving image quality.
- We design a distribution alignment initialization strategy to align the initial watermark encoding with the latent prior, thereby enabling earlier capture of the watermark signal and faster convergence.

Related Work

Diffusion Models

The DDPMs (Ho, Jain, and Abbeel 2020) have rapidly become one of the core directions in text-to-image generation research. Its process is based on Markov chains, and the generation of data is regarded as a gradual sampling and denoising process from noise, which can generate realistic and diverse images. Although DALLE-2 (Ramesh et al. 2022) and Imagen (Saharia et al. 2022) based on DDPMs demonstrated excellent synthesis quality, the high inference leads to high computational overhead. To balance image generation capability and efficiency, Rombach et al. (Rombach et al. 2022) proposed the Latent Diffusion Model, which first compresses high-dimensional images into a low-dimensional latent space using VAE. The perceptual compression stage preserved the overall structure of the image, while the generation stage focused on detail reconstruction. The combination of the two compensated for information loss, achieved high-resolution image synthesis, and reduced the computational complexity of diffusion model training and inference. After being trained on LAION-5B dataset (Schuhmann et al. 2022), this architecture became widely popular as Stable Diffusion and gave rise to a variety of high-quality generation schemes (Yun et al. 2025; Huang et al. 2025), accelerating the development of the AIGC field. However, its powerful generation capacity is also accompanied by security risks.

Watermarking Diffusion Models

In diffusion models, watermarking techniques have been integrated into multiple components to achieve deep fusion within the generation pipeline. At the decoder level, Fernandez et al. (Fernandez et al. 2023) proposed fine-tuning the VAE decoder so that the output image contained a 48-bit binary string, which was robust to image modification. Kim et al. (Kim et al. 2024) embedded 30-bit unique digital fingerprints into the decoder, making it difficult for end users to bypass or remove them. At the backbone network level, Feng et al. (Feng et al. 2024) and Wang et al. (Wang et al. 2025) both embedded 48-bit binary information by fine-tuning the U-Net. The former focused on the security of white-box scenarios, while the latter enhanced the resilience against downstream fine-tuning. At the latent space level, Yang et al. (Yang et al. 2024) proposed mapping 256-bit binary information into the latent space to achieve lossless watermark embedding. Wen et al. (Wen et al. 2023) introduced a scheme of embedding a tree-ring pattern in the latent frequency domain, which can be applied to any diffusion model. Zhang et al. (Zhang et al. 2025) designed a strategy to superimpose binary image watermarks containing metadata onto the latent space, achieving plug-and-play multi-variant compatibility.

Methodology

The Training Phase

The training process is described in three aspects: watermark embedding, watermark reconstruction, and loss function. To improve the training efficiency, the latent representation of original image is obtained via the VAE encoder, and only the sub-modules receive gradient update. The detailed overview of the Redundancy-Aware Latent Injection (RAIN) pipeline is presented in Figure 3.

Watermark Embedding Combining the inherent characteristics in latent space, the RGB image watermark is fused into the latent representation of the carrier image, achieving imperceptible watermark embedding. To solve the problem that watermark signals are difficult to learn early in training, we adopt a distribution alignment initialization strategy, using the pretrained VAE encoder $\mathcal{E}$ as the initial watermark encoder, aligning its output z_w with the latent prior I_w of the original image in distribution and scale. This strategy reduces initialization bias and enables the model to reach a stable optimization regime more quickly.

$$z_w = \mathcal{E}\big(I_w\big). \tag{1}$$

Subsequently, to mitigate the risk of disturbing highly sensitive regions and introducing artifacts or structural distortions due to forced watermark injection, we implement a redundancy-aware adaptive watermark fusion method. We first perform PCA on the flattened latent representation and redistribute each principal component's relative energy back to the original coordinates according to its squared projection, yielding the relative energy contribution $\tilde{E}_j$ at position j across all principal components:

$$\tilde{E}_j = \sum_{i=1}^{D} r_i \, u_{j,i}^2, \tag{2}$$

where r_i is the normalized explained-variance ratio of the i-th component, $\sum_i r_i = 1$, $D = 4 \times 64 \times 64$ is the dimensionality of the flattened latent, $\mathbf{u}_i \in \mathbb{R}^D$ is the corresponding unit eigenvector, $u_{j,i}$ is its component at the original coordinate j, and $\tilde{E}_j$ satisfies $\sum_{j=1}^{D} \tilde{E}_j = 1$.

Then, j is reshaped to (c, h, w) to obtain per-channel spatial energy contributions $\tilde{E}$. Following the principle "high energy implies low redundancy and conversely low energy implies high redundancy," we apply min–max normalization and take the complement to derive a differentiated redundancy distribution $\tilde{R}$:

$$\tilde{R} = 1 - \big(\frac{\tilde{E} - \min \tilde{E}}{\max \tilde{E} - \min \tilde{E}}\big), \tag{3}$$

The watermark latent z_w, cover latent z_{ori} and $\tilde{R}$ are used as inputs to the redundancy-guided fusion network $\mathcal{F}_{\mathcal{E}}$. By utilizing redundancy, the watermark is guided to adaptively redistribute and then fused with cover latent to obtain the watermarked latent z_{wm}, enabling focused embedding in more tolerant regions:

$$z_{wm} = \mathcal{F}_{\mathcal{E}}\big(z_{ori}, z_w, \tilde{R}\big). \tag{4}$$

By feeding the z_{wm} into the VAE decoder $\mathcal{D}$, the final watermarked image I_{wm} is reconstructed, thus completing the watermark embedding. This process not only enables efficient and stable injection of the high-capacity image watermark but also preserves the carrier image's visual quality.

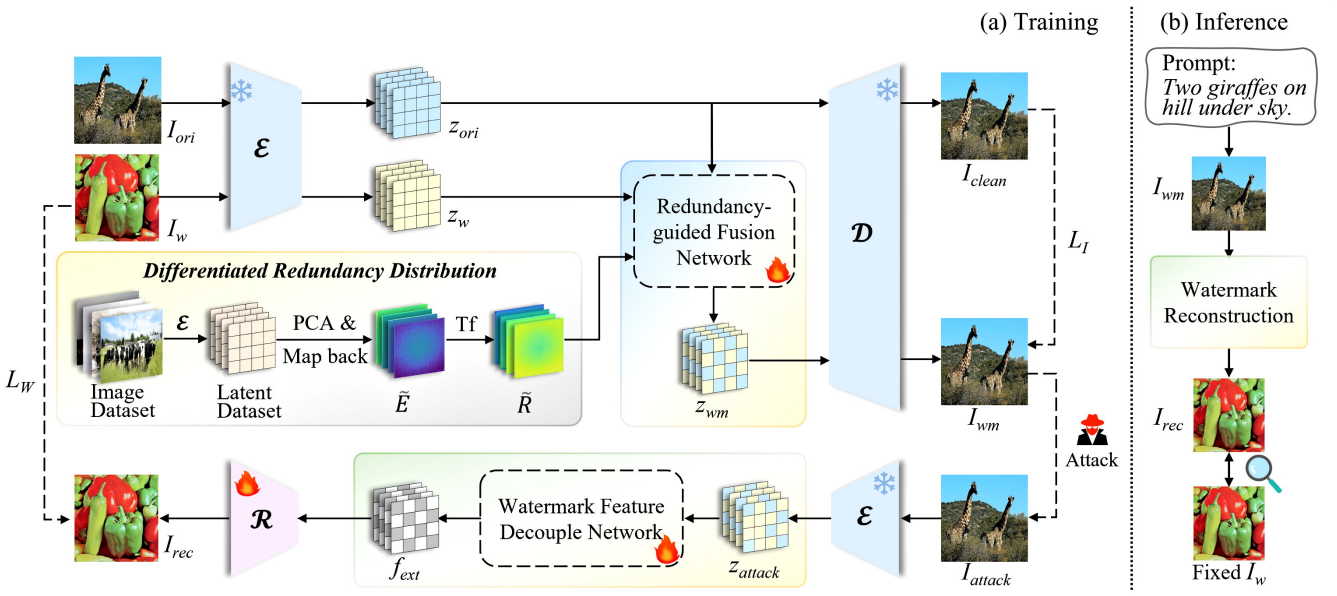


Figure 3: Overview of the RAIN pipeline. (a) Training phase: Sub-modules are trained to embed and extract the high-capacity image watermark guided by inherent redundancy distributions in the latent space. Tf stands for Transform. (b) Inference phase: For a suspect model, the watermark is extracted from its generated images to establish provenance of ownership. More details about network structure see in Appendix.

Watermark Reconstruction We recover the watermark in latent space for two reasons. On the one hand, during reconstruction from latent to image, the watermark signal is attenuated by decoding process, making direct extraction from the final image difficult. On the other hand, the latent preserves overall structure and semantics, so encoding a degraded image into latent space helps suppress pixel-level degradation caused by noise or compression. Consequently, the watermark hidden in the latent space exhibit more reliable recoverability. The I_{attack} is first encoded by $\mathcal{E}$ to obtain the corrupted latent z_{attack}. The watermark feature decouple network $\mathcal{F}_{\mathcal{D}}$ then selectively suppresses noise and non-watermark semantic components and preserves the subspace that carries the watermark payload to extracts the watermark feature f_{ext} from z_{attack}. Finally, the watermark reconstructor $\mathcal{R}$, built upon the VAE decoder architecture, decodes f_{ext} using its strong detail-restoration capability to reconstruct the watermark image:

$$I_{\text{rec}} = \mathcal{R}\big(\mathcal{F}_{\mathcal{D}}(\mathcal{E}(I_{\text{attack}}))\big). \tag{5}$$

At this stage, the watermark is decoupled and recovered from the latent representation. The recovered watermark I_{rec} is then compared with the original watermark to verify ownership of the generated content.

Loss Function The method aims to keep the watermarked image visually consistent with the cover and recover the watermark from watermarked image for traceability. Therefore, we have designed two loss functions L_I and L_W. L_I is used to measure the distance between the watermarked image I_{wm} and the clean image I_{clean}. L_W measures the distance between I_{rec} and the original watermark I_{w}, so the total loss function is the sum of these two loss functions:

$$L_{\text{total}} = L_I(I_{\text{wm}}, I_{\text{clean}}) + L_W(I_{\text{rec}}, I_{\text{w}}). \tag{6}$$

To ensure visual quality, we use both Mean Squared Error (MSE) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018):

$$\begin{aligned} L_I = \lambda_{MSE} \cdot MSE(I_{\text{wm}}, I_{\text{clean}}) \\ + \lambda_{LPIPS} \cdot LPIPS(I_{\text{wm}}, I_{\text{clean}}), \end{aligned} \tag{7}$$

where λ_{MSE} and λ_{LPIPS} are the weight coefficients. This combination optimizes visual consistency at both pixel and feature levels, ensuring image quality.

For watermark reconstruction, we use MSE to measure the difference between I_{rec} and I_{w}.

$$L_W = \lambda_{MSE_W} \cdot \text{MSE}(I_{\text{rec}}, I_{\text{w}}) \tag{8}$$

where λ_{MSE_W} ensure more concealed and stable watermark reconstruction. The overall goal is to minimize the total loss, enabling the joint optimization of image quality and watermark reconstruction performance.

The Inference Phase

The watermark embedder, watermark feature decoupler and watermark reconstructor, which have been trained and optimized, are seamlessly integrated into the inference process of the diffusion model in a plug-in manner. Following sampling and denoising, the watermark embedder automatically embeds the predefined watermark into the latent z_0, yielding the final watermarked output.

In scenarios such as suspected misuse, unauthorized redistribution, or provenance verification, it is necessary to

Watermark Form	Method	Capacity	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CLIPScore ↑
Binary bits	Stable Signature	48 bits	26.8112	0.8169	0.1083	26.2326 (+0.5294)	0.3183 (+0.0005)
	Aqualora	48 bits	17.9009	0.6739	0.2796	25.1753 (+0.5599)	0.3193 (–0.0005)
	SleeperMark	48 bits	16.1424	0.5680	0.3490	**23.8338** (–0.4189)	0.3091 (–0.0008)
	LW	64 bits	24.3566	0.7941	0.1508	25.1218 (-0.5814)	0.3183 (+0.0005)
	RoSteALS	100 bits	30.7722	0.9485	0.0580	29.7300 (+3.2809)	0.3289 (-0.0023)
Binary Image	MarkPlugger	256×256	37.0200	0.9430	0.0441	/ (-1.5000)	/
RGB Image	RAIN	512×512	**40.6300**	**0.9845**	**0.0199**	26.3712 (**-0.0226**)	**0.3313 (-0.0001)**
RGB Image	RAIN†	512×512	40.4696	0.9873	0.0140	26.3391 (-0.0547)	0.3315 (-0.0002)

Table 1: Evaluation of image visual quality and generation quality. For FID and CLIPScore, the difference with regards to original is shown in (·). Bold indicates the best result, † represents that the watermark is natural image (not included in Bold), otherwise the watermark is a QR code converted to RGB, and "/" means not available.

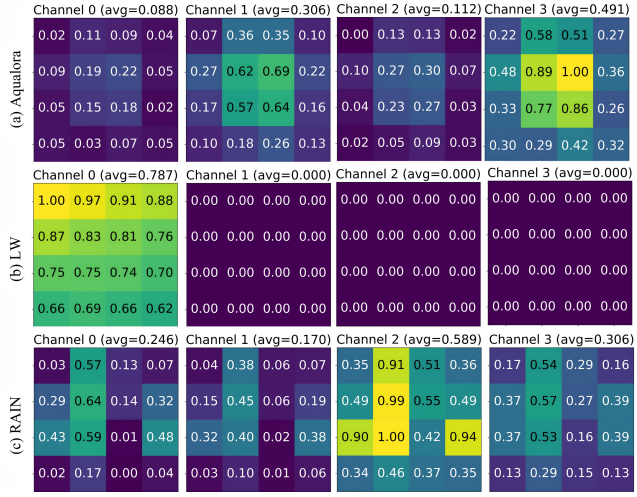


Figure 4: Watermark distribution heatmap: (a) AquaLoRA, (b) LW, and (c) RAIN.

trace the source. When given a suspect image, we first encode it into the latent space, then apply the watermark feature decoupler to extract the watermark representation, and finally the watermark reconstructor uses that representation to faithfully recover the watermark image. This plugin-based design requires no modification to the base network parameters, remains compatible with diverse latent diffusion variants, and provides a streamlined, dependable path for rapid deployment and version iteration.

Experiments

Experimental Settings

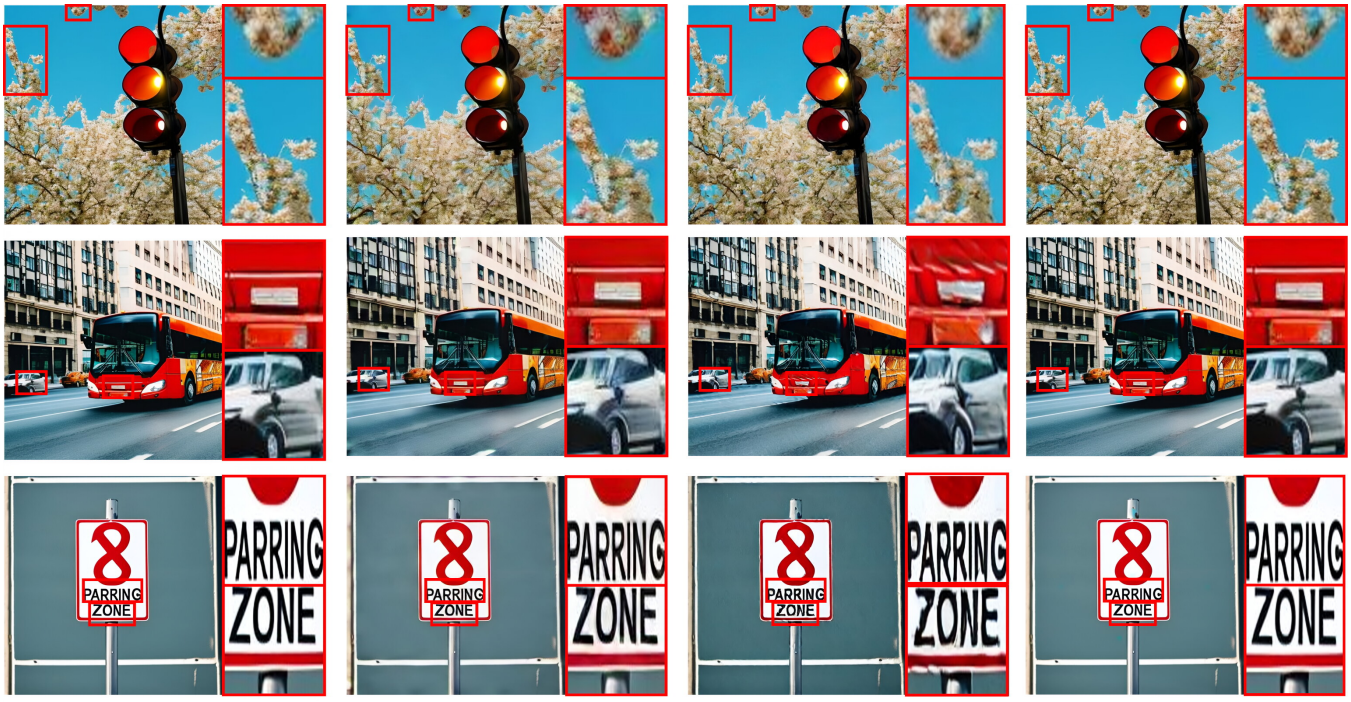
Datasets and models The COCO2017 dataset (Lin et al. 2014) serves as the primary data source: 10000 images for training, and 1000 images for validation. During inference, we generate and assess watermarked outputs using 2000 text prompts drawn from the COCO caption pool. The entire watermarking framework is implemented as external plug-ins on Stable Diffusion v2.1, enabling seamless integration without modifying the base model.

Evaluation Metrics To evaluate the efficacy and overall performance of the proposed method, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) (Hore and Ziou 2010) quantify pixel- and structure-level image consistency, LPIPS (Zhang et al. 2018) assesses perceptual alignment in deep feature space; Fréchet Inception Distance (FID) (Heusel et al. 2017) measures the distributional divergence between generated and real images, and the CLIP score (Radford et al. 2021) evaluates semantic alignment between generated content and text prompts. Watermark extraction accuracy (ACC) is used to assess the robustness under various perturbations, directly reflecting the attack resilience.

Implementation Details All experiments are implemented in PyTorch and executed on an single NVIDIA RTX 4090 GPU. Model parameters are optimized using the AdamW (Loshchilov and Hutter 2018) optimizer with a learning rate of 1e-4. The weight coefficients for the loss terms are set as λ_{MSE}=λ_{MSE_W}=1, and λ_{LPIPS}=3. Both the generated watermarked images and the watermark image are configured as 512×512 RGB format. In addition to the natural image as watermark, the QR-code watermark converted to RGB prior to encoding is also employed, facilitating direct comparison of extraction accuracy with related methods. The diffusion latent dimension is set to 4×64×64 in accordance with the Stable Diffusion v2.1.

Watermark Invisibility

Quantitative Evaluation To assess the impact of watermark embedding on both image visual quality and diffusion model generation performance, a set of evaluation metrics is employed, with results summarized in Table 1. At the pixel level, our PSNR reaches approximately 40 dB, and the SSIM is about 0.99, which significantly outperforming baseline methods. At the feature level, our LPIPS scores remain consistently lower than other approaches. For approaches that fine-tune diffusion model components, it is more ap-



Original Image RostaALS LW Ours

Figure 5: Visual comparison of the original image and watermarked images obtained by different methods. Each image's right side shows a zoomed-in view of the red-boxed region.

propriate to assess their impact on generative performance using FID and CLIPScore metrics. Other methods achieve better metric scores, whereas RAIN produces the smallest deviation from the original. The results demonstrate that our method can effectively achieve the high-capacity watermark embedding without reducing the generation quality. which is attributed to the full utilization of the differentiated redundancy distribution in the latent space.

To substantiate the situation, we construct and statistically analyze watermark distribution heatmaps for different methods as shown in Figure 4(a). Aqualora (Feng et al. 2024) uses a fixed watermark template, but the energy distribution is highly imbalanced and concentrated in channel 3. Some channels receive only weak signals, wasting capacity and robustness, while others are overloaded, risking perceptible artifacts. LW (Meng, Peng, and Dong 2025) injects the watermark exclusively into channel 0, resulting in large-scale modifications. This concentrated perturbation degrades image quality, and limits the overall watermark capacity due to other channels not being used. Their one-size-fits-all injection fails to consider to the latent space's redundancy and perceptual tolerance, creating trade-offs between capacity and visual quality In contrast, our method learns a better embedding pattern via the redundancy-aware adaptive watermark fusion method. By exploiting the latent space's differentiated redundancy and perceptual tolerance, it allocates stronger watermark signals to robust directions and minimizes disturbance in sensitive ones, preserving visual quality while achieving efficient and high-capacity embedding.

Qualitative Evaluation Furthermore, we conduct qualitative comparisons against baseline methods to visually assess watermark imperceptibility and image quality preservation. Representative examples generated by each approach under identical prompts are shown in Figure 5. The Comparison methods often introduce visible artifacts, such as blurring around object contours or texture degradation. Relatively speaking, our method preserves fine details and sharp edges, rendering the watermark effectively imperceptible to the human eye. These results demonstrate that RAIN maintains exceptional visual fidelity even when embedding high-capacity watermark information, highlighting the advantage of redundancy-aware guidance.

Watermarking Robustness

To comprehensively evaluate the robustness of each method in image degradation scenarios, the ACC is adopted as a unified metric. The results is summarized in Figure 6, covering common attacks such as JPEG (QF=70), Gaussian Blur (r=3), Resize (0.8), Gaussian Noise ($\mu = 0$, σ=0.01), Crop (10%), and Color Jitter. Except under the Crop attack, RAIN maintains a high ACC across the other five distortion types, matching or slightly surpassing existing methods. However, the Crop or other severe distortion attacks types cause more information loss and significantly reduces the resistance of RAIN, mainly due to the embedding of a large capacity RGB image watermark. Nevertheless, RAIN can still maintain a certain level of accuracy, demonstrating its tolerance for local information loss. Moreover, when the watermark is I_w

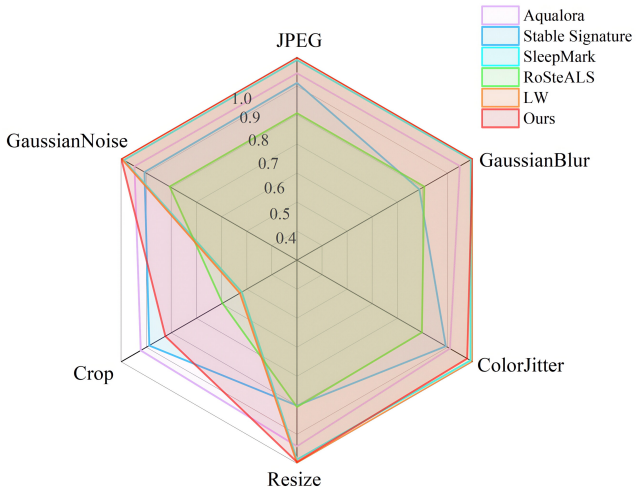


Figure 6: The robustness comparison of different methods.

in Figure 3, the PSNR for recovering the watermark reaches approximately 35 dB, which is sufficient for human perceptual recognition. Overall, the experimental results demonstrate that RAIN still has good cross-distortion robustness while embedding the high-capacity watermark and ensuring image quality, indicating that redundancy-aware watermark fusion method fully utilizes redundancy without completely sacrificing robustness.

Generalizability

Our method is plug-and-play within the diffusion generation pipeline and can be transferred directly to other LDM variants without retraining any existing components. After transfer, no significant degradation is observed on any key metric as shown in Tables 2. The results demonstrate robust cross-version generalization, which indicates that our method is highly adaptable to and compatible with other latent diffusion variants. In addition, RAIN utilizes the characteristics of latent space for watermark injection, so it can only be extended to models with latent space. More relevant analysis and results can be found in the Appendix.

Ablation Study

To validate the effectiveness of the redundancy-aware guidance and distribution alignment initialization, we conduct ablation experiments across four configurations as shown in Table 3. For the redundancy-aware guidance: compared to the Case 1, the use of guidance raises the initial imperceptibility and watermark recoverability at 2k iterations, with further gains at 10k iterations. This demonstrates that exploiting latent redundancy substantially improves image quality. For the distribution alignment initialization: the watermark recoverability at 2k iterations is only 14dB and ACC is 0 in Case 1. By contrast, applying distribution alignment in Case 3 boosts W_P to 35dB and ACC to 0.70, with these advantages persisting into later iterations. Moreover, the use of initialization reduces the required convergence steps from 25k to 13k, indicating that it not only accelerates convergence

Model	PSNR	SSIM	LPIPS	△FID	△CLIP
SD1-4	39.9330	0.9838	0.0200	-0.0213	0.0001
SD1-5	40.0096	0.9840	0.1990	0.0183	0.0001
SD2-0	40.5040	0.9842	0.0200	0.0135	-0.0002
SD2-1	40.6300	0.9845	0.0199	-0.0226	-0.0001

Table 2: Performance of our method across different Stable Diffusion variants. CLIP represents the metric CLIPScore.

Case	Redun	Align	Iter	I_P	W_P	ACC	C_Iter
1			2k	16.56	14.26	0.00	25k
			10k	20.44	35.57	0.70	
2	✓		2k	17.89	20.43	0.19	24k
			10k	24.20	35.94	0.73	
3		✓	2k	18.86	35.44	0.71	13k
			10k	23.88	41.58	0.84	
4	✓	✓	2k	20.34	37.11	0.79	14k
			10k	26.45	41.82	0.85	

Table 3: Impact of Strategy Configurations. Redun: Redundancy-aware guidance. Align: Distribution alignment initialization. Iter: Number of training iterations. I_P: Average PSNR (watermarked/clean images). W_P: Average PSNR (recovered/original watermarks). C_Iter: Approximate iterations required for initial convergence.

but also stabilizes early watermark quality. They complement each other: the watermark signal is captured quickly and stably to accelerate convergence, while embedding focuses on more tolerant regions to preserve image quality.

Conclusion

To address the shortcoming of existing methods that neglect inherent redundancy and perceptual characteristics when embedding watermarks in the latent space, this work presents the Redundancy-Aware Latent Injection (RAIN) framework. We systematically quantify the redundancy distribution within the latent space to guide differentiated watermark embedding, and develop a distribution alignment initialization strategy. The former uses the redundancy heatmap to adaptively allocate embedding amplitude by applying heavier payloads in high-tolerance regions and reducing perturbations in low-redundancy regions, thereby achieving low-distortion watermark injection. The latter uses a pretrained VAE encoder to align the initial watermark encoding with the latent prior, enabling early capture of the watermark signal to speed up convergence and boost embedding capacity. Extensive quantitative and qualitative results demonstrate that RAIN preserves exceptional image quality while embedding the high-capacity RGB image watermark.

Acknowledgements

This work was supported in part by the National NSF of China (No.U24B20179, No.62336001, No.62120106009), the Talent Fund of Beijing Jiaotong University (No. 2024XKRC011), Beijing Natural Science Foundation (No. 4264127), and China Postdoctoral Science Foundation(Grant No. 2025M781453).

References

Baluja, S. 2020. Hiding Images within Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7): 1685–1697.

Ding, G.; Yang, C.; Wang, S.; Li, X.; Zhang, J.; Jin, X.; and Huang, Q. 2025. Dis^{2}Booth: Learning Image Distribution with Disentangled Features for Text-to-Image Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2744–2752.

Fang, H.; Chen, K.; Yang, Z.; Cui, B.; Zhang, W.; and Chang, E.-C. 2025. CoSDA: Enhancing the Robustness of Inversion-based Generative Image Watermarking Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2888–2896.

Feng, W.; Zhou, W.; He, J.; Zhang, J.; Wei, T.; Li, G.; Zhang, T.; Zhang, W.; and Yu, N. 2024. AquaLoRA: Toward White-box Protection for Customized Stable Diffusion Models via Watermark LoRA. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Fernandez, P.; Couairon, G.; Jégou, H.; Douze, M.; and Furon, T. 2023. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22466–22477.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 6840–6851.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2366–2369.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6): 417.

Huang, K.; Duan, C.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2025. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47: 3563–3579.

Ke, J.; Wong, W.; Wang, J.; Li, M.; Fei, L.; and Wen, J. 2025. DiffusionREC: Diffusion Model with Adaptive Condition for Referring Expression Comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 4221–4229.

Kim, C.; Min, K.; Patel, M.; Cheng, S.; and Yang, Y. 2024. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8974–8983.

Li, Z.; Song, Y.; Tao, R.; Jia, X.; Zhao, Y.; and Wang, W. 2025. Unsupervised region-based image editing of denoising diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 18638–18646.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 740–755. Springer.

Loshchilov, I.; and Hutter, F. 2018. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Masrani, V.; Akbari, M.; Yue, D. M. X.; Rezaei, A.; and Zhang, Y. 2025. Task-Agnostic Language Model Watermarking via High Entropy Passthrough Layers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 24849–24857.

Meng, Z.; Peng, B.; and Dong, J. 2025. Latent Watermark: Inject and Detect Watermarks in Latent Diffusion Space. *IEEE Transactions on Multimedia*, 27: 3399–3410.

Panaitescu-Liess, M.-A.; Che, Z.; An, B.; Xu, Y.; Pathmanathan, P.; Chakraborty, S.; Zhu, S.; Goldstein, T.; and Huang, F. 2025. Can watermarking large language models prevent copyrighted text generation and hide training data? In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 25002–25009.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763. PmLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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 (CVPR)*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 25278–25294.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024a. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 5052–5060.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28130–28139.

Wang, Z.; Guo, J.; Zhu, J.; Li, Y.; Huang, H.; Chen, M.; and Tu, Z. 2025. Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8213–8224.

Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 58047–58063.

Xian, X.; Wang, G.; Bi, X.; Srinivasa, J.; Kundu, A.; Hong, M.; and Ding, J. 2024. Raw: A robust and agile plug-and-play watermark framework for ai-generated images with provable guarantees. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 132077–132105.

Yang, Z.; Zeng, K.; Chen, K.; Fang, H.; Zhang, W.; and Yu, N. 2024. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12162–12171.

Yun, T.; Zhang, D.; Park, J.; and Pan, L. 2025. Learning to sample effective and diverse prompts for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23625–23635.

Zhai, R.; Ni, R.; Chen, Y.; Yu, Y.; and Zhao, Y. 2023. Defending fake via warning: Universal proactive defense against face manipulation. *IEEE Signal Processing Letters*, 30: 1072–1076.

Zhang, G.; Wang, L.; Su, Y.; and Liu, A.-A. 2025. MarkPlugger: Generalizable Watermark Framework for Latent Diffusion Models without Retraining. *IEEE Transactions on Multimedia*, 1–9.

Zhang, J.; Chen, D.; Liao, J.; Ma, Z.; Fang, H.; Zhang, W.; Feng, H.; Hua, G.; and Yu, N. 2024. Robust model watermarking for image processing networks via structure consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10): 6985–6992.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 586–595.