GenPTW: Latent Image Watermarking for Provenance Tracing and Tamper Localization

Zhenliang Gan[1], Chunya Liu[2], Yichao Tang[1], Binghao Wang[2], Shiwen Cui[2], Weiqiang Wang[2], Xinpeng Zhang[1]*

[1]College of Computer Science and Artificial Intelligence, Fudan University, China
[2]Ant Group, China
zlgan23@m.fudan.edu.cn, {liuchunya.lcy, weiqiang.wwq, binghao.wbh, donn.csw}@antgroup.com,
{yichao_tang, zhangxinpeng}@fudan.edu.cn

Abstract

The proliferation of generative image models has revolutionized AIGC creation while amplifying concerns over content provenance and manipulation forensics. Existing methods are typically either unable to localize tampering or restricted to specific generative settings, limiting their practical utility. We propose **GenPTW**, a **Gen**eral watermarking framework that unifies **P**rovenance tracing and **T**amper localization in latent space. It supports both in-generation and post-generation embedding without altering the generative process, and is plug-and-play compatible with latent diffusion models (LDMs) and visual autoregressive (VAR) models. To achieve precise provenance tracing and tamper localization, we embed the watermark using two complementary mechanisms: cross-attention fusion aligned with latent semantics and spatial fusion providing explicit spatial guidance for edit sensitivity. A tamper-aware extractor jointly conducts provenance tracing and tamper localization by leveraging watermark features together with high-frequency features. Experiments show that GenPTW maintains high visual fidelity and strong robustness against diverse AIGC-editing.

Introduction

Generative models are advancing at an unprecedented pace, particularly text-to-image diffusion models such as Stable Diffusion, DALL·E 3, and Imagen, as well as the emerging VAR models that demonstrate remarkable potential. These models are capable of synthesizing highly realistic and visually striking images with flexible editability, reshaping the paradigm of visual content creation and dissemination. However, such impressive generative and editing capabilities also pose a double-edged sword, introducing security risks such as content misuse, ambiguous ownership, and challenges in tamper detection.These risks underscore two core challenges: attributing content ownership and detecting potential manipulations.

Image watermarking has been widely adopted for source tracing and copyright protection. However, most existing methods remain centered on ownership identification and offer limited support for tamper localization. Accurate localization is essential for distinguishing generated content from subsequent edits, assigning responsibility to the originating model, and understanding the intent behind manipulation. These functions together underpin a comprehensive framework for forensic traceability. Recent studies have begun to unify provenance tracing and tamper localization. For example, *SepMark* (Wu, Liao, and Ou 2023) introduces a separable watermark architecture for robust ownership verification and deepfake detection, while *EditGuard* (Zhang et al. 2024) exploits spatial fragility in steganography to identify tampered regions. A common limitation of these approaches is their reliance on post-generation embedding, where watermarks are added only after image synthesis. This design decouples watermarking from the generative process, complicates deployment, and reduces overall efficiency.

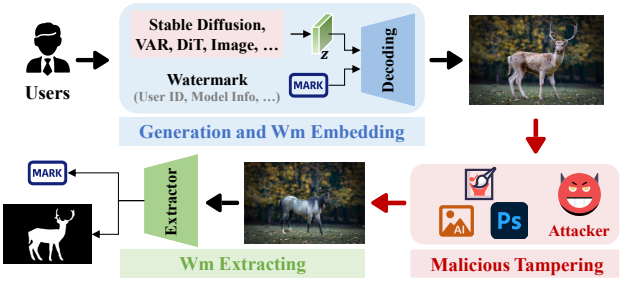

Figure 1: The process of embedding and extracting **GenPTW** for dual forensic objectives.

To address these limitations, recent research has shifted toward an *in-generation* paradigm, injecting watermarks directly during image generation. Despite this advancement, most existing approaches still lack tamper localization capabilities and remain fragile under downstream AIGC editing, thus limiting their forensic reliability and real-world applicability. Moreover, current designs are largely confined to diffusion models, lacking a unified watermarking framework that can generalize across LDMs, VAR models, and post hoc scenarios involving pre-generated images. Most image generation models, including LDMs and VAR models, operate in the latent space and decode to images afterward. To address limitations in forensic capacity and model general-

*Corresponding Author.

ization, we propose GenPTW, a general-purpose image watermarking framework that embeds watermarks in the latent space. GenPTW preserves the original generation pipeline and is compatible with LDMs, VAR, and post hoc processing of existing images. As shown in Fig. 1, GenPTW targets a dual forensic objective: **(1) Provenance tracing**: Identifying the source and ownership of images. **(2) Tamper Localization:** Marking the regions that have been manipulated.

To minimally affect the generative process while enabling precise tamper localization, we introduce two dedicated modules. First, the Cross-Attention Fusion (CAF) module dynamically selects optimal watermark embedding strategies conditioned on latent features, balancing robustness and imperceptibility. Second, the Spatial Fusion (SF) module expands the watermark into a spatial guidance map and injects it into the final upsampled latent features of the decoder to enhance tamper localization. A gradient-guided encoder further embeds the watermark under Just Noticeable Difference (JND) constraints, guided by a modification cost map and regularized across multi-scale latent features. Finally, a Tamper-Aware Extractor tightly integrates provenance tracing and tamper localization through a unified feature backbone, ensuring effective and robust decoding. Our main contributions are summarized as follows:

- We propose **GenPTW**, the first framework to unify provenance tracing and tamper localization through a single embedding in latent space, seamlessly supporting both in-generation and post-hoc usage.
- We design a novel embedding strategy employing Cross-Attention and Spatial Fusion modules to achieve adaptive watermarking, that improves imperceptibility of watermark, robustness of watermark extraction, and precision of tampering localization.
- We introduce a tamper-aware extractor that jointly leverages embedded watermark cues and high-frequency features to enable robust watermark decoding and accurate tamper localization, even under severe degradations.
- Extensive experiments show that GenPTW achieves superior performance over existing watermarking and forensic baselines in terms of visual fidelity, flexibility, and robustness.

Related Work

Image Tamper Localization

Localization of image tampering is a critical task in digital media forensics, mainly categorized into passive and proactive methods. ***Passive methods*** examine intrinsic attributes such as statistical features, lighting conditions, color distribution, noise discrepancies, and DCT correlations (Chen et al. 2021; Wu, Abd-Almageed, and Natarajan 2018; Islam et al. 2020; Zhuang et al. 2021; Guillaro et al. 2023; Yu et al. 2024) to identify tampering without external information. But they require real domain data for optimal performance and lack generalization ability. ***Proactive methods*** involve embedding imperceptible markers or watermarks into images, which are easily destroyed or altered when tampering occurs. Traditional fragile watermarking method, such as block-wise hash verification or pixel-level grayscale analysis (Cheng, Ni, and Zhao 2012; Lin et al. 2023; NR and Shreelekshmi 2022; Hurrah et al. 2019), have limited localization accuracy and flexibility. To address these limitations, deep learning-based approaches(Wang et al. 2021; Asnani et al. 2023) have been developed. More recently, methods like EditGuard (Zhang et al. 2024) and OmniGuard (Zhang et al. 2025) have employed two-stage embedding for pixel-level localization and copyright protection, through combining steganography and watermark technology. However, they still require Pre-defined template to ensure precise pixel-level tamper localization.

Post-hoc Image Watermarking

Post-hoc watermarking methods embed watermarks into already existing images, and encompass both traditional and deep learning-based approaches. Traditional methods design embedding mechanisms in imperceptible spatial(Chan and Cheng 2004) or frequency domains(Navas et al. 2008) to insert watermarks. Deep learning methods like Hidden (Zhu et al. 2018) and CIN (Ma et al. 2022) employ encoder-noise-decoder architectures to learn robust watermarking schemes, while techniques like MBRS (Jia, Fang, and Zhang 2021), StegaStamp (Tancik, Mildenhall, and Ng 2020), Pimog (Fang et al. 2022), LFM (Wengrowski and Dana 2019), and DeNol (Fang et al. 2023) utilize differentiable noise layers during training to simulate real-world distortions—such as JPEG compression, screenshot capture, or photographic degradation—to enhance the robustness of the embedded watermark. Furthermore, such as Robust-Wide (Hu et al. 2024) designed denoise sampling guidance module and OmniGuard (Zhang et al. 2025) proposed a lightweight AIGC editing simulation layer to against AIGC-editing.

In-Generation Image Watermarking

In-generation watermarking refers to embedding watermarks directly during the image creation process, rather than via post-processing. Previous works such as WatermarkDM (Zhao et al. 2023), ProMark (Asnani et al. 2024), and Diffusion-Shield (Cui et al. 2023) have embedded copyright watermarks into training datasets to influence dataset attribution, allowing extraction of the watermark from images generated by models trained or fine-tuned on these datasets. Obviously, this technique is inflexible and resource-intensive.

Recent research has identified two primary strategies: ***Initial Noise Modulation***. Methods like Tree-Ring (Wen et al. 2023), GaussMarker (Li et al. 2025) and Gaussian Shading (Yang et al. 2024) embed watermark features by modifying the Fourier spectrum of initial Gaussian noise vectors, or inject encrypted Gaussian-distributed patterns into the initial noise. However, altering the randomness of initial noise distributions can negatively impact both the quality and diversity of generated images, since these changes may reduce the natural appearance and variety of outputs. ***Latent Space Adaptation***. For instance, Stable Signature(Fernandez et al. 2023) fine-tune the VAE decoder for each digital fingerprint, which limits scalability. WOUAF(Kim et al. 2024) and FSW(Xiong et al. 2023) embed flexible watermarks by

introducing auxiliary networks and fine-tuning the VAE decoder. RoSteALS (Bui et al. 2023) explored the feasibility of leveraging latent space redundancy to embed watermarks without modifying the decoder. Similarly, LaWa (Rezaei et al. 2024) and WMAdapter(Ci et al. 2024) integrate watermark features into latent variables via auxiliary networks while keeping decoder parameters frozen, thus maintaining scalability and efficiency.

The above methods only meet the requirements for watermark extraction. In this paper, GenPTW achieves both watermark extraction and tamper location through single-stage embedding, while maintaining high fidelity and robustness.

Method

Overall Framework of GenPTW

As shown in Fig. 2, GenPTW embeds watermark information into the latent space to achieve provenance tracing and tamper localization within a unified design. Given a binary message, GenPTW encodes it and injects the signal into the latent decoding process without altering the original generative model. CAF adaptively selects embedding strategies based on latent semantics, while SF injects spatial priors into the final decoding layer to support fine-grained localization. In the extraction phase, the image is processed by a shared encoder to extract features for both watermark decoding and tamper localization. The shared backbone encourages consistency between tasks and improves joint performance. To enhance robustness against AIGC edits, we incorporate a distortion simulation layer during training. In addition, a JND-aware perceptual loss constrains watermark perturbations through a pixel-wise cost map, effectively preserving visual fidelity. Each component of the framework is detailed in the following subsections.

Multi-scale Latent Space Embedding

In general, diffusion models are trained either in the image space or in a compact latent space to reduce computational cost and memory consumption. For LDMs, the output of the diffusion process lies in this latent space. Similarly, VAR models also operate in a latent space to improve modeling efficiency and scalability. These models typically incorporate an image autoencoder to map images into a compact latent space, where the image $I_{\text{source}} \in R^{H\times W\times 3}$ is encoded into a latent representation $z = \mathcal{E}(I_{\text{source}}) \in R^{H/\alpha\times W/\alpha\times C}$ by a factor of α, usually $\alpha = 8$, and decoded by $\mathcal{D}(z)$ in a multi-stage manner. During generation, the latent representation $\mathbf{Z}_{\text{latent}}$ can be obtained in three different ways: it may be synthesized by a diffusion model, generated by a VAR model, or produced through image compression. The decoder then progressively upsamples $\mathbf{Z}_{\text{latent}}$ to reconstruct the final image.

To embed watermark information, we adopt a coarse-to-fine strategy that injects the message into latent features at multiple decoder stages. The decoder performs $\log_2(\alpha)$ upsampling operations to recover the original image resolution, and we embed a scale-specific watermark after each upsampling step. Given a k-bit binary watermark message $Wm \in \{0,1\}^k$, a watermark encoder E_{wm} generates the initial watermark feature f_{w_1} that matches the shape of z_{latent}. This feature is added to the latent before the first upsampling stage. We expect f_{w_1} to evolve alongside the decoding path and progressively inject the watermarking.

Cross-Attention Fusion Module To facilitate the gradual refinement of watermark features during decoding, we employ a Cross-Attention Fusion (CAF) module at each intermediate decoding stage $i \in 1,\ldots,\log_2(\alpha)-1$. As illustrated in Fig. 3, the latent representation z_i and the evolving watermark feature f_{w_i} are conceptualized as two interacting modalities. Unlike prior approaches that directly superimpose watermark signals onto feature maps, our design allows the latent representation to actively guide the injection of watermark information. The CAF module generates a residual watermark feature that aligns with the current latent distribution, enabling structurally coherent integration throughout the decoding process—much like ink naturally diffusing along the texture of paper:

$$f_{w_{i+1}} = CAF_i(f_{w_i}, z_i) \tag{1}$$

CAF Structure. The CAF block implements a cross-attention mechanism in which f_{w_i} serves as the query, and z_i is used as both the key and value. The inputs are first projected into a shared embedding space via 1×1 convolutions:

$$Q = \text{Conv}_Q(f_{w_i}),\ K = \text{Conv}_K(z_i),\ V = \text{Conv}_V(z_i) \tag{2}$$

The attention output is computed using scaled dot-product attention:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{3}$$

The resulting context feature is concatenated with f_{w_i} and passed through a lightweight residual convolutional block to obtain the updated watermark feature:

$$f_{w_{i+1}} = \text{Conv}_{\text{Res}}\left([\text{Attention}(Q,K,V),\ f_{w_i}]\right) \tag{4}$$

Finally, the updated watermark feature is added to the current latent representation to form the watermarked latent:

$$z_{wm_i} = z_i + f_{w_{i+1}} \tag{5}$$

The fused latent z_{wm_i} is then forwarded to the next decoding stage, allowing the watermark signal to evolve coherently alongside the image features. This interaction mechanism ensures that watermark information is integrated in a latent-aware and transformation-consistent manner, without disrupting the generative pathway.

Spatial Fusion Module At the final stage ($i = \log_2(\alpha)$), the latent representation reaches the full spatial resolution of the image. At this point, most structural and textural information has been reconstructed through the decoder, and the final latent features serve as the immediate precursor to the output image. To inject dense and spatially aligned watermark information at this high-resolution level, we introduce a Spatial Fusion (SF) module. It transforms the initial watermark feature f_{w_1} into a spatial prior and injects it into the final latent representation via feature-wise residual fusion:

$$z_{\text{wm}_i} = \text{SF}(f_{w_1}, z_i) \tag{6}$$

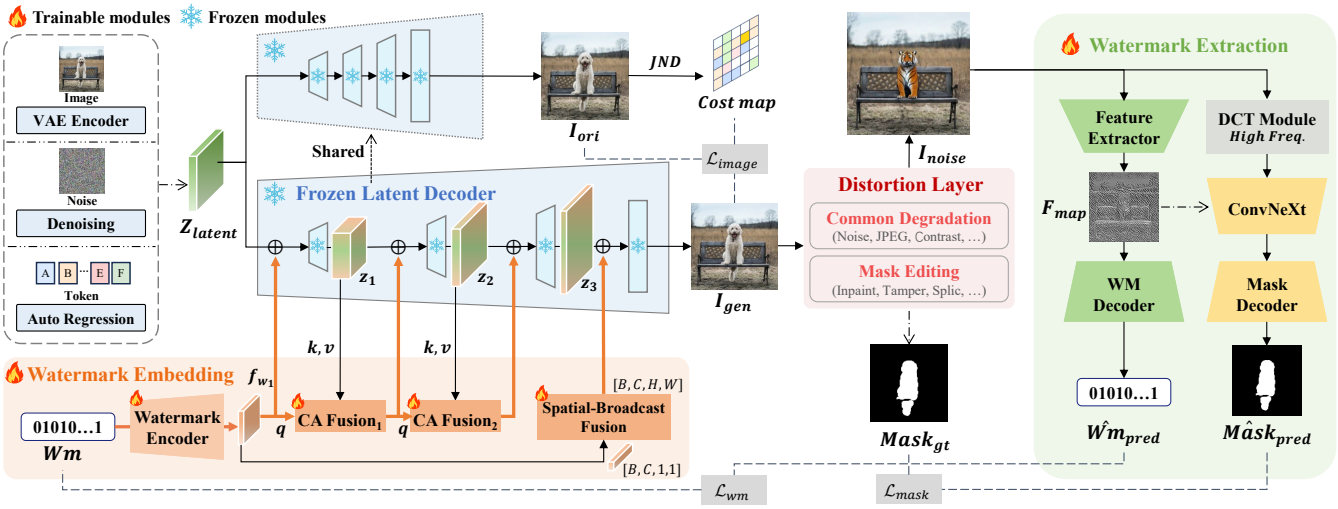

Figure 2: The Framework of **GenPTW**. A Wm plug-in inserted during generation **without modifying the original model**.

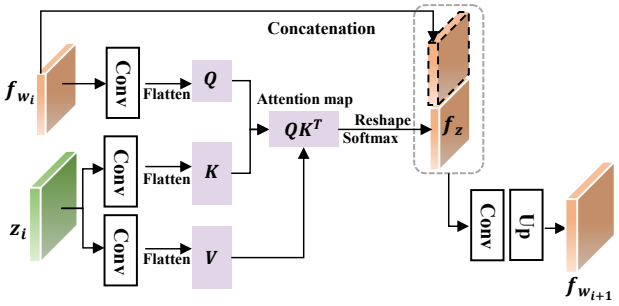

Figure 3: Illustration of Cross-Attention Fusion Block.

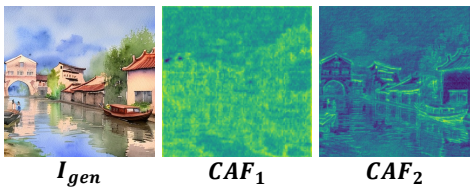

Figure 4: Average attention maps during image generation with the watermark module.

SF Structure. The SF module first flattens the initial watermark feature $f_{w_1} \in R^{B\times C'\times H'\times W'}$ into a global vector $\phi \in R^{B\times D}$, which is then projected to match the channel dimension of the target latent feature through a linear layer:

$$\phi' = \text{Linear}(\text{Flatten}(f_{w_1})) \in R^{B\times C} \tag{7}$$

This projected vector ϕ' is broadcast spatially to match the shape of the full-resolution latent feature $z \in R^{B\times C\times H\times W}$:

$$\Phi = \text{Broadcast}(\phi') \in R^{B\times C\times H\times W} \tag{8}$$

The broadcasted watermark prior Φ is concatenated with the latent feature z along the channel dimension and passed through a lightweight convolutional fusion block. The result is integrated back into the latent space via residual addition:

$$z_{\text{wm}_i} = z + \text{Conv}_{\text{fuse}}(\text{Concat}(z_i, \Phi)) \tag{9}$$

This spatial injection mechanism enables the watermark signal to be embedded explicitly at each pixel location, making the final representation more responsive to localized perturbations. By aligning a globally consistent watermark with the full-resolution latent features, the SF module facilitates fine-grained tamper detection without compromising image reconstruction fidelity.

Finally, the remaining decoding layers convert the watermarked latent representation into the generated image $\mathbf{I}_{\text{gen}}$. This output is then passed through a distortion layer to produce the degraded version $\mathbf{I}_{\text{noise}}$. Detailed configurations are provided in the Appendix.

Tamper-Aware Extractor

To enable robust watermark decoding and accurate tamper localization in a cooperative manner, we propose a Tamper-Aware Extractor that jointly exploits embedded watermark cues and high-frequency features for manipulation analysis. Unlike prior designs that treat watermark extraction and tamper detection as independent tasks, our framework structurally integrates the two objectives into a unified pipeline where they mutually reinforce each other.

Specifically, we employ a shared feature extractor E_f to process the distorted image $\mathbf{I}_{\text{noise}}$ and generate a shared feature representation $\mathbf{F}_{\text{map}}$. This feature map is then passed to the watermark decoder D_{wm} to recover the embedded watermark message:

$$\mathbf{F}_{\text{map}} = E_f(\mathbf{I}_{\text{noise}}),\ \hat{\mathbf{Wm}}_{\mathbf{pred}} = D_{\text{wm}}(\mathbf{F}_{\text{map}}) \tag{10}$$

To improve localization sensitivity to subtle structural inconsistencies, we extract high-frequency priors from $\mathbf{I}_{\text{noise}}$ using the Discrete Cosine Transform (DCT) (Gonzalez and

Woods 2018), yielding a high-frequency map $\mathbf{I}_h$ that highlights potential tampering artifacts.

The feature map $\mathbf{F}_{\text{map}}$ is concatenated with $\mathbf{I}_h$ to form the combined input $\mathbf{I}_{\text{all}}$, which is then fed into a ConvNeXt-based multi-scale encoder CN_{Enc} to extract hierarchical semantic features:

$$\mathbf{I}_{\text{all}} = \{\mathbf{I}_h,\ \mathbf{F}_{\text{map}}\} \tag{11}$$

$$\begin{aligned} &\{F_{s_1}, F_{s_2}, F_{s_3}, F_{s_4}\} = CN_{\text{Enc}}(\mathbf{I}_{\text{all}}), \\ &F_{s_i} \in R^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i},\ i = 1, 2, 3, 4. \end{aligned} \tag{12}$$

Here, C_i denotes the number of channels at each scale. The multi-level features are subsequently fused and decoded to yield a dense tampering probability map.

To this end, we adopt a hierarchical mask decoder D_{mask} that first upsamples each feature map to a common resolution using transposed convolutions. The aligned features are then concatenated and processed through a lightweight convolutional head to predict the final tampering mask:

$$\hat{\mathbf{Mask}}_{pred} = D_{\text{mask}}(\{F_{s_1}, F_{s_2}, F_{s_3}, F_{s_4}\}) \tag{13}$$

The performance of watermark extraction is measured using binary cross-entropy loss between the predicted watermark $\hat{\mathbf{Wm}}_{pred}$ and the ground-truth message $\mathbf{Wm}$:

$$\mathcal{L}_{wm} = \lambda_k\, \ell_{\text{bce}}(\hat{\mathbf{Wm}}_{pred}, \mathbf{Wm}) \tag{14}$$

For tamper localization, we compute a combination of binary cross-entropy loss and edge-aware loss (Ma et al. 2023) between the predicted mask $\hat{\mathbf{Mask}}_{pred}$ and the ground-truth mask $\mathbf{Mask}_{gt}$:

$$\begin{aligned} \mathcal{L}_{mask} = \lambda_m\, \ell_{\text{bce}}(\hat{\mathbf{Mask}}_{pred},\, \mathbf{Mask}_{gt}) \\ + \gamma\, \ell_{\text{edge}}(\hat{\mathbf{Mask}}_{pred},\, \mathbf{Mask}_{gt}) \end{aligned} \tag{15}$$

where γ is set to 20.

Ensuring Visual Quality

Our method jointly embeds watermark signals for provenance tracing and tamper localization, which inevitably introduces stronger perturbations than single-task designs. To preserve visual fidelity, we incorporate a Just-Noticeable-Difference (JND)–guided loss to suppress perceptible artifacts. During training, we generate a clean image $\mathbf{I}_{\text{ori}}$ (without watermark injection) and a watermarked image $\mathbf{I}_{\text{gen}}$ (with injection enabled). The clean image is only used for supervision and perceptual loss computation.

We first compute a map $\text{JND}(\mathbf{I}_{\text{ori}}) \in R^{3\times H\times W}$ to estimate pixel-level visibility thresholds, and construct a cost map:

$$\text{Cost Map} = 1 - \alpha_{\text{JND}} \cdot \text{JND}(\mathbf{I}_{\text{ori}}) \tag{16}$$

The JND-weighted residual loss is defined as:

$$\ell_{ct} = \text{Cost Map} \odot \mathbf{I}_{\text{gen}} \tag{17}$$

To further constrain distortion, we combine pixel-wise MSE:

$$\ell_I = \|\mathbf{I}_{\text{gen}} - \mathbf{I}_{\text{ori}}\|_2^2 \tag{18}$$

with the LPIPS loss (Zhang et al. 2018), which better captures perceptual differences. The total visual quality loss is:

$$\mathcal{L}_{\text{image}} = \lambda_I \ell_{\text{I}} + \lambda_{LPIPS} \ell_{LPIPS} + \lambda_{ct} \ell_{ct} \tag{19}$$

	SD Inp.		Splicing		Lama	
Method	F1	AUC	F1	AUC	F1	AUC
Post-hoc / COCO Test Images						
MVSS-Net	0.182	0.491	0.419	0.802	0.027	0.508
CAT-Net	0.148	0.676	0.200	0.722	0.149	0.727
PSCC-Net	0.170	0.504	0.192	0.688	0.134	0.333
IML-ViT	0.215	0.592	0.469	0.759	0.107	0.460
EditGuard	0.964	0.973	0.932	0.989	0.967	0.966
OmniGuard	0.960	**0.997**	0.916	**0.995**	0.951	**0.999**
GenPTW*	0.946	**0.997**	0.930	0.994	0.956	0.998
In-generation / AI-generated						
GenPTW	**0.973**	0.996	**0.965**	0.993	**0.976**	0.997

Table 1: Localization performance of GenPTW and SOTA baselines under clean conditions. GenPTW* is post-hoc; GenPTW is in-generation (SD-based).

Experiments

Experimental Setup

We train our models on the MS COCO dataset (Lin et al. 2014), using segmentation annotations to generate mask. All images and masks are resized to 512×512, and editing prompts are fixed as "None". The test set includes 5,000 natural images from the COCO validation set and 1,000 AI-generated images synthesized by Stable Diffusion v2 from COCO captions. Editing prompts and masks follow the UltraEdit protocol (Zhao et al. 2024). For VAR-based models, we evaluate 1,000 images, one per class from 0 to 999, each paired with a randomly generated mask. Since our watermark is embedded in latent space, pairing it with a corresponding autoencoder enables post-hoc usage. We therefore evaluate three pipelines: Stable Diffusion-based generation, VAR-based generation, and a post-hoc variant. Training is conducted on an NVIDIA A100 GPU using the AdamW optimizer (learning rate 1×10^{-5}, batch size 2, gradient accumulation steps 8) with a cosine learning rate schedule. Additional implementation details are provided in the Appendix.

Comparision with Localization Methods

To evaluate the tamper localization performance of our proposed *GenPTW*, we compare it against several state-of-the-art passive detection methods, including PSCC-Net (Liu et al. 2022), MVSS-Net (Dong et al. 2022), CAT-Net (Kwon et al. 2021), and IML-ViT (Ma et al. 2023), as well as proactive watermark-based approaches EditGuard (Zhang et al. 2024) and OmniGuard (Zhang et al. 2025). We adopt F1 score and AUC as evaluation metrics, with evaluation conducted on the aforementioned test set. Manipulations include Stable Diffusion Inpaint (Rombach et al. 2022), the Lama (Suvorov et al. 2022), and classical splicing, covering both AIGC-based and traditional manipulations.

Figure 6 compares tamper localization results across methods. Passive approaches like PSCC-Net and IML-ViT often miss manipulations under complex edits, while proactive methods such as EditGuard tend to produce noisy or incomplete masks, with performance sensitive to hyperparameters. In contrast, GenPTW consistently yields accurate and

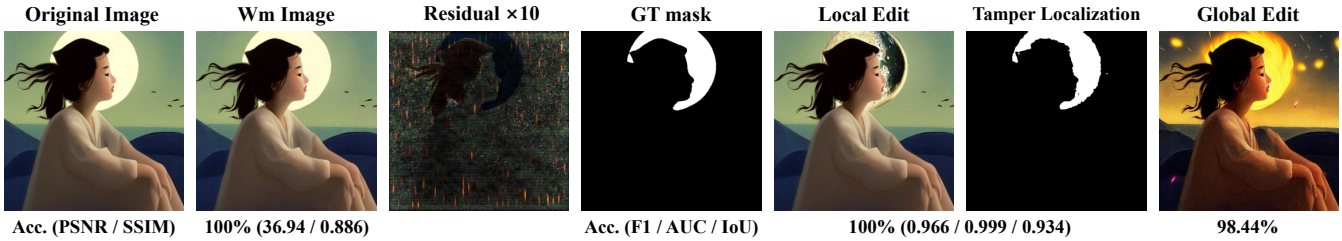

Figure 5: Qualitative examples of generated images using GenPTW.

			Global Edit		Local Edit			Common Degradation			
Method	PSNR/SSIM	LPIPS/SIFID	P2P	SDIP+	SDIP	Lama	RD	Blur	Contr	Bright	JPEG
Post-hoc Watermarking on COCO Test Images											
PIMoG(30)	35.94/0.891	0.098/0.045	0.656	0.618	0.913	0.949	0.954	0.725	0.962	0.934	0.936
SepMark(30)	31.95/0.879	0.112/0.054	0.876	0.908	0.951	0.952	0.948	0.949	**0.987**	0.955	0.996
EditGuard(64)	37.12/0.902	0.082/0.012	0.534	0.596	0.969	0.971	0.960	0.723	0.950	**0.984**	0.938
OmniGuard(100)	37.21/**0.912**	**0.069/0.009**	0.934	**0.970**	**0.997**	**0.994**	**0.995**	**0.980**	**0.987**	0.960	**0.998**
Robust-Wide(64)	**39.18**/0.905	0.097/0.044	**0.976**	0.956	**0.997**	0.968	0.981	0.787	0.976	0.968	0.981
WOUAF*(64)	28.51/0.791	0.144/0.067	0.543	0.564	0.825	0.819	0.863	0.951	0.928	0.943	0.962
Lawa*(48)	32.94/0.813	0.096/0.023	0.537	0.584	0.851	0.833	0.879	0.973	**1.000**	**1.000**	**0.998**
GenPTW*(64)	**35.56/0.864**	**0.077/0.002**	**0.963**	**0.975**	**0.998**	**0.997**	**0.996**	**0.999**	**1.000**	**1.000**	0.996
In-generation Watermarking on AI-Generated											
Stable Signature(48)	31.43/0.834	0.123/0.064	0.561	0.626	0.905	0.894	0.884	0.784	0.914	0.943	0.914
WOUAF(64)	30.71/0.847	0.130/0.061	0.587	0.601	0.874	0.883	0.916	0.981	0.971	0.975	0.991
Lawa(48)	35.14/0.821	0.073/0.033	0.591	0.629	0.892	0.897	0.926	0.999	**1.000**	**1.000**	0.998
GenPTW(64)	**39.56/0.892**	**0.069/0.002**	**0.969**	**0.974**	**0.990**	**0.994**	**0.977**	**1.000**	**1.000**	0.998	**1.000**

Table 2: **Fidelity and bit accuracy of GenPTW vs. SOTA baselines.** * denotes post-hoc. + denotes image regeneration via an inpainting model. RD denotes random cropping on $\mathbf{I}_{\text{gen}}$, with cropped regions replaced by $\mathbf{I}_{\text{ori}}$.

well-aligned masks across diverse perturbations, without requiring post-processing or parameter tuning. For full-image semantic rewriting tasks like InstructP2P(Brooks, Holynski, and Efros 2023), GenPTW remains effective in watermark extraction and tamper detection. Since such edits fundamentally alter global structure, the model often flags the entire image as tampered—reflecting a deliberate design choice to prioritize preserving original visual semantics over adapting to aggressive content shifts.

Comparison with Deep Watermarking

We comprehensively evaluate the performance of GenPTW against both in-generation and post-generation watermarking methods, including Stable Signature, WOUAF, and LaWa for in-generation, and PIMoG (Fang et al. 2022), SepMark (Wu, Liao, and Ou 2023), EditGuard (Zhang et al. 2024), OmniGuard (Zhang et al. 2025), and Robust-Wide (Hu et al. 2024) for post-generation. All methods are evaluated on 512×512 images from the UltraEdit dataset, except for the 256-resolution VAR-based variant. Test settings follow those described in the experimental setup.

As shown in Table 2, *GenPTW* consistently achieves higher bit recovery accuracy under various perturbations, with a PSNR of **39.56**dB—surpassing all in-generation baselines and outperforming several post-generation methods. Figure 5 presents examples generated by Stable Diffusion v2, followed by both local (SD Inpaint) and global (InstructP2P) edits. Despite significant changes in visual style and structure, GenPTW successfully recovers the embedded watermark, demonstrating strong robustness and generalizability to both global and localized real-world manipulations. Furthermore, we evaluate GenPTW by training separate models on VAR (Tian et al. 2024), DiT (Peebles and Xie 2023), and SDXL (Podell et al. 2023) architectures. As shown in Table 3, the results demonstrate its generalizability across diverse generative models.

Method	ACC	PSNR	SSIM	F1	AUC
VAR(256)	0.961	27.56	0.814	0.945	0.991
VAR(512)	0.985	35.26	0.873	**0.989**	**0.999**
DIT(512)	**0.989**	**38.48**	**0.902**	0.974	0.996
SDXL(1024)	0.981	38.01	0.897	0.957	0.995

Table 3: Evaluation on different models under SD Inpaint.

Ablation Study

Effect of Watermark Embedding Modules We perform ablations by progressively removing different embedding modules. As shown in Table 4, the SF module has a significant impact on tamper localization, demonstrating its role in enhancing spatial alignment and local sensitivity at high resolution. Removing the CAF_2 module leads to a substantial drop in tamper localization performance, while remov-

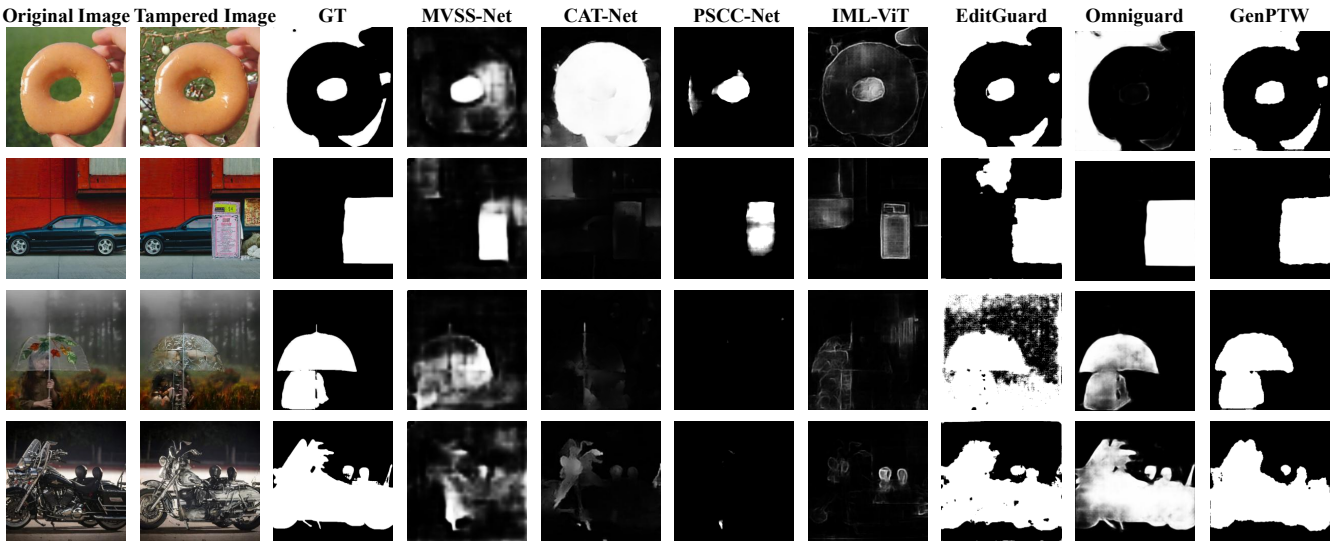

Figure 6: Visualized Comparison between our GenPTW and other methods.

CAF_1	CAF_2	SF	ACC.	PSNR	F1
-	-	-	0.594	38.87	0.520
✓	-	-	0.986	**40.87**	0.721
✓	✓	-	0.987	39.22	0.901
✓	✓	✓	**0.990**	39.56	**0.973**

Table 4: Ablation of Embedding Modules under SD Inpaint.

Method	ACC.	PSNR	SSIM	F1	AUC
ADD	**0.997**	30.06	0.805	0.743	0.596
Fuse	0.989	18.00	0.369	0.845	0.739
CAF	**0.997**	**35.54**	**0.863**	**0.861**	**0.981**

Table 5: Ablation Study of the CAF under SD Inpaint.

ing CAF_1 primarily affects watermark extraction accuracy. These results indicate that hierarchical embedding is essential for robust watermark propagation.

Effect of CAF Modules To evaluate the effectiveness of the CAF, we ablate its cross-attention structure and replace the watermark injection with either direct addition (ADD) or convolutional fusion (Fuse). As shown in Table 5, the ADD variant exhibits poor robustness, while Fuse enhances tamper localization at the expense of visual fidelity. In contrast, the complete CAT module achieves an optimal balance between robustness and perceptual quality. Fig. 4 shows that the two CAFs attend to different structure–texture boundary information; when combined with the evidence in Table 4, it indicates that CAF_1 enhances watermark robustness whereas CAF_2 improves localization accuracy. We provide the overhead analysis in the appendix.

Impact of Tamper Localization Inputs We conduct an ablation study on the input configurations of the tamper localization branch, as shown in Table 6. Using either the high-

$\mathbf{I}_{\text{noise}}$	$\mathbf{I}_h$	$\mathbf{F}_{\text{map}}$	ACC.	PSNR	F1
✓	-	-	0.964	35.59	0.958
-	✓	-	0.970	36.82	0.959
✓	-	✓	**0.999**	36.77	0.962
-	✓	✓	0.991	**37.85**	**0.974**

Table 6: Impact of different input configurations for the tamper localization branch under SD Inpaint.

frequency image $\mathbf{I}_h$ or $\mathbf{I}_{\text{noise}}$ alone achieves acceptable performance, while incorporating the fused feature $\mathbf{F}_{\text{map}}$ significantly improves both tamper localization and watermark extraction. The $\mathbf{I}_h + \mathbf{F}_{\text{map}}$ combination achieves accuracy comparable to $\mathbf{I}_{\text{noise}} + \mathbf{F}_{\text{map}}$ while providing the best visual quality, yielding the optimal trade-off.

Conclusion

In this paper, we propose ***GenPTW***, a **Gen**eral latent-space watermarking framework for proactive **P**rovenance tracing and **T**amper localization. To the best of our knowledge, GenPTW is the first framework to unify these two capabilities through a single embedding in latent space, seamlessly supporting both in-generation and post-hoc usage. To achieve this, We propose a novel strategy that adapts the watermark to image latent space, incorporating both a cross-attention fusion module and a spatial fusion module. The cross-attention fusion module embeds the watermark based on latent features, while the spatial fusion module integrates the watermark as spatial cues into final features. Furthermore, we introduce a tamper-aware extractor that combines watermark signals with high-frequency features to enable precise tamper localization. Extensive experiments demonstrate that GenPTW consistently outperforms SOTA watermarking and forensic baselines in fidelity, localization accuracy, and robustness across diverse manipulation scenarios.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 62450067, U22B2047 and 62502093). The authors from Ant Group are supported by the Leading Innovative and Entrepreneur Team Introduction Program of Hangzhou (Grant No. TD2022005).

References

Asnani, V.; Collomosse, J.; Bui, T.; Liu, X.; and Agarwal, S. 2024. ProMark: Proactive diffusion watermarking for causal attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10802–10811.

Asnani, V.; Yin, X.; Hassner, T.; and Liu, X. 2023. Malp: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Bui, T.; Agarwal, S.; Yu, N.; and Collomosse, J. 2023. RoSteALS: Robust Steganography using Autoencoder Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 933–942.

Chan, C.-K.; and Cheng, L.-M. 2004. Hiding data in images by simple LSB substitution. *Pattern recognition*.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Cheng, B.; Ni, R.; and Zhao, Y. 2012. A refining localization watermarking for image tamper detection and recovery. In *2012 IEEE 11th International Conference on Signal Processing*, volume 2, 984–988.

Ci, H.; Song, Y.; Yang, P.; Xie, J.; and Shou, M. Z. 2024. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*.

Cui, Y.; Ren, J.; Xu, H.; He, P.; Liu, H.; Sun, L.; and Tang, J. 2023. DiffusionShield: A Watermark for Copyright Protection against Generative Diffusion Models. *arXiv preprint arXiv:2306.04642*.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Fang, H.; Chen, K.; Qiu, Y.; Liu, J.; Xu, K.; Fang, C.; Zhang, W.; and Chang, E.-C. 2023. Denol: a few-shot-sample-based decoupling noise layer for cross-channel watermarking robustness. In *Proceedings of the 31st ACM international conference on multimedia*, 7345–7353.

Fang, H.; Jia, Z.; Ma, Z.; Chang, E.-C.; and Zhang, W. 2022. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*.

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

Gonzalez, R. C.; and Woods, R. E. 2018. *Digital Image Processing*. Pearson, 4th edition.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, R.; Zhang, J.; Xu, T.; Li, J.; and Zhang, T. 2024. Robustwide: Robust watermarking against instruction-driven image editing. In *European Conference on Computer Vision*, 20–37. Springer.

Hurrah, N. N.; Parah, S. A.; Loan, N. A.; Sheikh, J. A.; Elhoseny, M.; and Muhammad, K. 2019. Dual watermarking framework for privacy protection and content authentication of multimedia. *Future generation computer Systems*, 94: 654–673.

Islam, A.; Long, C.; Basharat, A.; and Hoogs, A. 2020. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia, Z.; Fang, H.; and Zhang, W. 2021. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM International Conference on Multimedia (MM)*.

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*, 8974–8983.

Kwon, M.-J.; Yu, I.-J.; Nam, S.-H.; and Lee, H.-K. 2021. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Li, K.; Huang, Z.; Hou, X.; and Hong, C. 2025. GaussMarker: Robust Dual-Domain Watermark for Diffusion Models. *arXiv preprint arXiv:2506.11444*.

Lin, C.-C.; Lee, T.-L.; Chang, Y.-F.; Shiu, P.-F.; and Zhang, B. 2023. Fragile Watermarking for Tamper Localization and Self-Recovery Based on AMBTC and VQ. *Electronics*, 12(2): 415.

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)*.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.

Ma, R.; Guo, M.; Hou, Y.; Yang, F.; Li, Y.; Jia, H.; and Xie, X. 2022. Towards Blind Watermarking: Combining Invertible and Non-invertible Mechanisms. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

Ma, X.; Du, B.; Jiang, Z.; Hammadi, A. Y. A.; and Zhou, J. 2023. IML-ViT: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*.

Navas, K.; Ajay, M. C.; Lekshmi, M.; Archana, T. S.; and Sasikumar, M. 2008. Dwt-dct-svd based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*, 271–274. IEEE.

NR, N. R.; and Shreelekshmi, R. 2022. Fragile watermarking scheme for tamper localization in images using logistic map and singular value decomposition. *Journal of Visual Communication and Image Representation*, 85: 103500.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Rezaei, A.; Akbari, M.; Alvar, S. R.; Fatemi, A.; and Zhang, Y. 2024. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, 118–136. Springer.

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)*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2117–2126.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37: 84839–84865.

Wang, R.; Juefei-Xu, F.; Luo, M.; Liu, Y.; and Wang, L. 2021. Faketagger: Robust safeguards against deepfake dissemination via provenance tracking. In *Proceedings of the 29th ACM international conference on multimedia*, 3546–3555.

Wen, Y.; Kirchenbauer, J.; Geiping, J.; and Goldstein, T. 2023. Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. *arXiv preprint arXiv:2305.20030*.

Wengrowski, E.; and Dana, K. 2019. Light field messaging with deep photographic steganography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1515–1524.

Wu, X.; Liao, X.; and Ou, B. 2023. SepMark: Deep Separable Watermarking for Unified Source Tracing and Deepfake Detection. In *Proceedings of the ACM international conference on Multimedia (MM)*.

Wu, Y.; Abd-Almageed, W.; and Natarajan, P. 2018. Image copy-move forgery detection via an end-to-end deep neural network. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1907–1915. IEEE.

Xiong, C.; Qin, C.; Feng, G.; and Zhang, X. 2023. Flexible and secure watermarking for latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1668–1676.

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*, 12162–12171.

Yu, Z.; Ni, J.; Lin, Y.; Deng, H.; and Li, B. 2024. DiffForensics: Leveraging Diffusion Prior to Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12765–12774.

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 conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Li, R.; Yu, J.; Xu, Y.; Li, W.; and Zhang, J. 2024. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11964–11974.

Zhang, X.; Tang, Z.; Xu, Z.; Li, R.; Xu, Y.; Chen, B.; Gao, F.; and Zhang, J. 2025. OmniGuard: Hybrid Manipulation Localization via Augmented Versatile Deep Image Watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3008–3018.

Zhao, H.; Ma, X.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024. UltraEdit: Instruction-based Fine-Grained Image Editing at Scale. arXiv:2407.05282.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Cheung, N.-M.; and Lin, M. 2023. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *European Conference on Computer Vision (ECCV)*.

Zhuang, P.; Li, H.; Tan, S.; Li, B.; and Huang, J. 2021. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security*, 16: 2986–2999.