

APT: Anchor-aligned Perturbations for Tamper Localization in Fully Regenerated Images

Suhyeon Ha, Woo Jae Kim, Joonsung Jeon, Sooel Son^{*}, and Sung-eui Yoon^{*}

KAIST

{suhyeon.ha, wkim97, mikeraph, sl.son, sungeui}@kaist.ac.kr
<https://suhyeonha.github.io/APT>

Abstract. Proactive tamper localization embeds an imperceptible signal into an image prior to distribution, enabling pixel-level manipulation detection. Existing methods assume a spliced (SP) setting, where synthesized regions are composited onto the original background, leaving embedded signals intact. However, real-world diffusion-based inpainting operates in a fully regenerated (FR) setting, where the entire image undergoes denoising, disrupting background signals and rendering existing frameworks ineffective. We propose APT, a semi-fragile latent-space perturbation that embeds a dense, vector-wise localization signal. By aligning each spatial feature vector toward a fixed anchor direction, APT localizes tampering via the alignment disparity between synthesized foreground and anchor-aligned background features after inpainting. The proposed hard negative mining loss and noisy perturbation branch further enforce uniform alignment. Experiments on COCO demonstrate that APT achieves an FR IoU of 0.92, outperforming the strongest baseline (WAM, 0.84), while existing methods collapse to near-random performance ($\text{AUC} \approx 0.5$), establishing APT as a practical forensic framework generalizable across tampering types unknown at test time.

Keywords: Tamper Localization · Image Inpainting · Proactive Perturbation

1 Introduction

Recent advances in diffusion-based inpainting [12, 19, 22, 30] enable highly photorealistic, text-guided image manipulation. These generative capabilities substantially lower the technical barrier for fraudsters, facilitating the large-scale production and dissemination of deceptive visual forgeries on social media, particularly during high-stakes global crises [2, 26]. Such misuse transforms inpainting from a creative tool into an instrument for information manipulation. To counter this threat, proactive tamper localization has emerged as a defensive scheme. The core idea is to embed an imperceptible signal into an image prior to distribution. Upon suspected manipulation, verifiers can analyze the protected image and localize the altered regions.

^{*} Co-corresponding authors

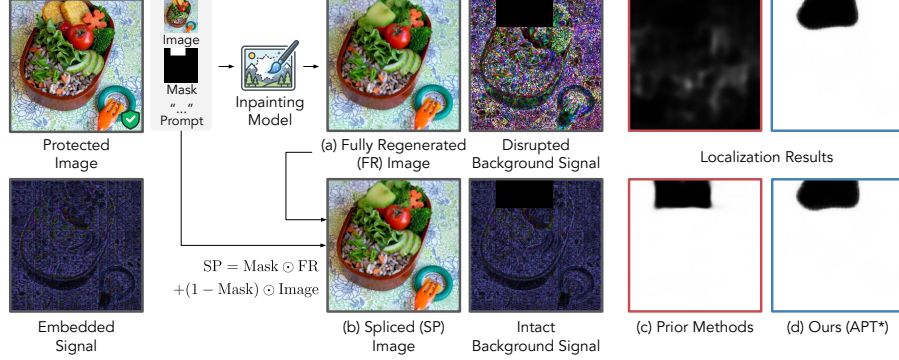


Fig. 1: Vulnerability of proactive localization under fully regenerated (FR) inpainting. In the spliced (SP) setting (b), assumed by existing methods, synthesized regions are composited onto the original background, leaving the embedded signal intact. In contrast, the FR setting (a) uses the raw inpainting output directly, disrupting the background signal across the entire image. While prior methods (c) struggle to localize tampered regions under FR due to severe background signal disruption, our method (d) produces accurate localization masks under both FR and SP settings. Background signals are visualized as $10 \times (|tamper - original| \odot (1 - mask))$.

Existing proactive localization methods [24, 28, 31, 32] rely on an optimistic adversary model that assumes splicing (SP): after synthesizing a forged region, the adversary copies the forged patch back onto the original image as a post-processing step (Fig. 1). Under this assumption, the background remains unchanged and the embedded signal outside the edited region is preserved, enabling reliable localization. However, this assumption is broken by diffusion-based editing in its native form. The raw output of diffusion models is a fully regenerated (FR) image; the entire image undergoes the noise sampling and denoising process to blend the synthesized foreground with the background. Accordingly, the embedded signal in the background is also impaired, even when the semantic content outside the edited region appears intact. While existing studies [8, 20] note that FR images are generally harder to detect than SP images, no prior work in proactive localization has addressed this practical and challenging FR setting.

Addressing this gap requires rethinking what properties a localization signal must satisfy to be effective in the FR setting. Diffusion-based inpainting models synthesize the target foreground based on a text prompt while reconstructing the original background. Successful localization under such generative processes thus requires a dual property: the signal must remain robust to background regeneration yet fragile to foreground synthesis.

However, existing proactive methods [24, 28, 32] struggle to satisfy this requirement, particularly the robustness to background regeneration, due to two primary limitations. First, their localization embeddings are fragile to both the VAE encoding-decoding process and pixel regeneration used in diffusion pipelines. As a result, signals embedded in background regions become unsta-

ble, rendering them unreliable as localization cues in the FR setting. Second, they often rely on global embeddings that lack explicit spatial correspondence between the signal and its location, making it difficult to accurately map pixel-level tampering.

To address these challenges, we propose APT, a semi-fragile perturbation framework that embeds a robust vector-wise localization signal in the VAE latent space. APT optimizes a latent perturbation that drives each spatial latent vector toward a predefined anchor vector, producing vector-wise alignment at every spatial location rather than relying on a single global embedding. Since we sample this anchor vector from the unit hypersphere independently of the generative model’s learned distribution, feature vectors synthesized by the diffusion model naturally deviate from this anchor direction. In contrast, during diffusion-based inpainting, background regions are reconstructed by heavily conditioning on the original latent representation, which already carries the anchor-induced alignment. This foreground-background alignment disparity yields a reliable localization signal that differentiates tampered and protected regions after inpainting. To further stabilize the signal, we introduce a hard negative mining loss that suppresses weakly aligned spatial features and a noisy perturbation branch that improves robustness of the embedded signal against VAE-induced distortions during background regeneration.

APT achieves consistent localization performance across both spliced and fully regenerated tampered images. At verification, the vector-wise detection map is directly upsampled to produce a pixel-level mask (APT), while an optional lightweight decoder can be employed to further refine spatial boundaries (APT*).

We compare APT with state-of-the-art tamper localization methods across both SP and FR settings. While existing methods perform well in SP settings, they suffer severe performance drops in IoU ranging from 0.14 to 0.93 when faced with FR manipulations. Specifically, StableGuard [28] and OmniGuard [32] collapse to near-random AUC levels (≈ 0.5), whereas APT maintains a robust AUC of 0.97. Furthermore, APT* surpasses the top-performing baseline (WAM [24], 0.84), with an IoU of 0.92, effectively bridging the gap in FR tamper localization.

Our contributions are as follows:

- We present the first systematic evaluation of proactive tamper localization under fully regenerated (FR) inpainting. We reveal a critical vulnerability in existing methods, suffering severe performance degradation on FR forgeries.
- We propose APT, a novel semi-fragile perturbation framework that optimizes a dense, vector-wise anchor signal in the VAE latent space, where the foreground-background alignment disparity serves as the localization cue. A hard negative mining loss and a noisy perturbation branch further enforce spatially uniform alignment, enabling robust localization across both SP and FR manipulations.
- Experiments on COCO demonstrate that APT significantly outperforms state-of-the-art baselines under the challenging FR manipulations while maintaining competitive performance on SP images. This establishes APT as a practical tamper localization framework for real-world scenarios where the manipulation type is unknown at test time.

2 Related Work

2.1 Diffusion-based Inpainting

Diffusion-based inpainting [12, 17, 19, 22, 30] takes as input an image, a binary mask, and a text prompt, and synthesizes content within the masked region so that the generated foreground is semantically consistent with the prompt and visually coherent with the surrounding image. To preserve background consistency, recent inpainting models integrate spatial context into the denoising process through several mechanisms. For example, [3, 17, 19] merge the denoised foreground with the noised original background according to the provided mask. [19, 22] concatenate the mask, the masked background image, and the noisy latent along the channel dimension and feed them as inputs to the U-Net [23] in the inpainting models. Also, [12, 30] inject mask-conditioned features from a secondary branch into a frozen U-Net feature through zero convolutions.

With these background-preserving mechanisms, modern inpainting models generate visually coherent outputs without requiring any post-processing. Consequently, adversaries are able to distribute the raw outputs of these models, in which the entire images are regenerated through the diffusion process. Therefore, tamper localization remains a critical challenge especially when synthesized images do not involve explicit splicing—a setting we refer to as fully regenerated (FR).

2.2 Tamper Localization

Traditional passive localization methods [5, 9–11, 15, 29] identify tampered regions by analyzing multi-view boundary artifacts [5], extracting hierarchical fine-grained features [10], detecting noise anomalies [9], or leveraging large multi-modal models [11]. While effective against manual editing techniques (*e.g.* copy-move, splicing), these passive methods struggle to detect photorealistic manipulations produced by advanced generative models [31, 32].

To address the limitations of passive methods, recent methods proactively embed localization signals before alterations. MaLP [1] learns proactive spatial templates by incorporating a generative model directly into the framework. EditGuard [31] utilizes invertible networks in wavelet domain to hide a secret image, while WAM [24] treats watermarking as a localized segmentation task. OmniGuard [32] extends EditGuard by injecting VAE compression and degradation-aware extractor. Furthermore, StableGuard [28] integrates watermarking during latent diffusion generation, utilizing a mixture-of-experts network for forensic extraction.

Despite recent advances in proactive tamper localization, existing methods struggle to generalize to FR inpainting due to two primary limitations. First, methods such as WAM [24] and OmniGuard [32] embed signals in the pixel or frequency domain, making them inherently susceptible to the VAE bottleneck and background regeneration. Second, methods such as StableGuard [28] apply

a single global signal across the entire image, lacking a direct spatial correspondence between the signal and its location. Without locally dense signals, mask predictors tend to overfit to visual cues such as splicing boundaries and collapse entirely from the global distortion under FR settings. This gap motivates a framework that combines latent-space perturbation optimization with spatially dense signal embedding.

3 Preliminary

3.1 Spliced vs. Fully Regenerated Images

The key difference between spliced (SP) and fully regenerated (FR) manipulations lies in whether the unmasked background is preserved. Let $x' = \mathcal{G}(x, m, c)$ denote the output of the generative model $\mathcal{G}(\cdot)$ conditioned on image x , mask m ($m = 1$ for the editing region), and prompt c . For an SP image, x'_{SP} is composited onto the original background using the binary mask m :

$$x'_{\text{SP}} = x' \odot m + x \odot (1 - m), \quad (1)$$

where $\odot$ denotes the element-wise product. In this case, the original background is strictly preserved, ensuring $x'_{\text{SP}} \odot (1 - m) = x \odot (1 - m)$. Consequently, any proactive signals embedded in the background remain undisturbed. In FR, the raw output from the generative model is used directly:

$$x'_{\text{FR}} = x'. \quad (2)$$

Here, the entire image undergoes stochastic noise injection and iterative denoising. Even regions outside the masked area are re-synthesized. Thus, pixel-level equivalence no longer holds (*i.e.* $x_{\text{FR}} \odot (1 - m) \neq x \odot (1 - m)$).

This discrepancy arises from multiple factors: (1) statistical disharmony between the replaced background and the model’s own background generation, (2) generative priors that regularize pixel values toward the learned data distribution [17], and (3) latent reconstruction errors in VAE-based inpainting models, which prevent exact recovery of original pixel-level details [12, 22]. Therefore, existing proactive localization methods become vulnerable to adversaries who employ diffusion-native FR, necessitating a detection framework that remains robust under full-image regeneration.

3.2 Anchor-aligned Proactive Perturbation

To establish a proactive defense, we build upon the zero-bit watermarking framework introduced in [7], which determines the presence of a hidden mark. Given an original image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, an imperceptible perturbation is added to generate a marked image $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$. The perturbation is optimized via gradient descent to push an image feature vector $\mathbf{v}$ toward a predefined anchor vector $\mathbf{a}$. This process effectively constrains the image feature within a decision boundary

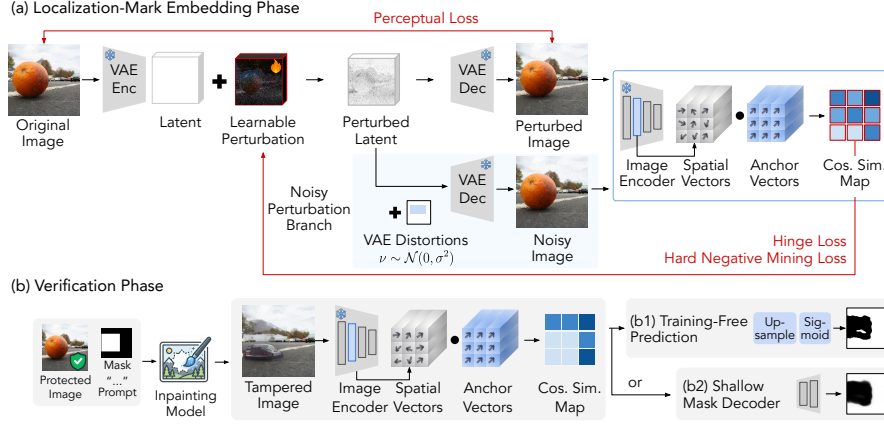


Fig. 2: Overview of the APT pipeline. (a) In the localization-mark embedding phase, a learnable perturbation is optimized in the VAE latent space to align spatial feature vectors toward predefined anchors. (b) In the verification phase, a per-vector cosine similarity map extracted from a potentially tampered image is passed to either (b1) a training-free predictor or (b2) a shallow mask decoder to produce a localization mask.

known as a hypercone $\mathcal{C}$, which is formally defined as the set of vectors satisfying:

$$\mathcal{C} = \left\{ \mathbf{v} \mid \frac{\mathbf{v} \cdot \mathbf{a}}{\|\mathbf{v}\| \|\mathbf{a}\|} \geq \cos(\theta) \right\}, \quad (3)$$

where the feature is geometrically aligned within an angular distance θ from the anchor $\mathbf{a}$. During detection, an image is identified as marked if its extracted feature vector resides within $\mathcal{C}$.

4 Method

We propose APT, a semi-fragile proactive localization framework that embeds a dense, vector-wise signal via anchor alignment in latent space (Sec. 4.1). APT embeds an imperceptible perturbation into an input image such that, in the resulting protected image, background features remain consistently aligned with predefined anchors even after full regeneration (FR).

Overview. Fig. 2 illustrates the overall pipeline of APT, which consists of two phases: localization-mark embedding and verification. In the localization-mark embedding phase, APT begins with an input image under protection and predefined anchor vectors. The perturbed latent, along with noisy variants of the perturbation, is decoded through the VAE decoder to produce two perturbed image variants. These images are subsequently processed by an image encoder that outputs spatial feature vectors. APT optimizes the learnable perturbation to align these spatial features toward predefined anchor directions. The key insight is that anchor-aligned background features persist—even under FR—whereas

synthesized foreground regions follow the generative model’s native distribution and exhibit low alignment with the anchor. This distributional separation enables precise, pixel-level localization without assuming strict background preservation. For this, we propose (1) a hard negative mining loss and (2) a noisy perturbation branch that enable spatially uniform alignment (Sec. 4.2). We also propose training-free and shallow mask detectors tailored to SP and FR forgeries (Sec. 4.3) that do not require any tampered or optimized samples.

The verification phase takes a potentially tampered image as input, extracts per-vector anchor alignment maps, and produces a pixel-level localization mask via either training-free prediction or a shallow mask decoder.

4.1 Anchor-aligned Perturbation for Localization

Embedding vector-wise perturbation. Given an original image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, APT directly optimizes the perturbation in the VAE latent space to enhance the robustness against VAE compression and pixel regeneration. Given an initial latent $\mathbf{z} = \mathcal{E}_{\text{vae}}(\mathbf{x})$ extracted from the original image using a pretrained VAE encoder $\mathcal{E}_{\text{vae}}$, we add a learnable latent perturbation $\boldsymbol{\delta}$ to construct the perturbed latent $\tilde{\mathbf{z}} = \mathbf{z} + \boldsymbol{\delta}$. The perturbed latent $\tilde{\mathbf{z}}$ is decoded via $\mathcal{D}_{\text{vae}}$ to obtain a temporary perturbed image $\tilde{\mathbf{x}}_{\text{perturb}} = \mathcal{D}_{\text{vae}}(\tilde{\mathbf{z}})$, which is passed through $\mathcal{E}_{\text{img}}$ to extract a dense feature map $\mathbf{F} \in \mathbb{R}^{H' \times W' \times d}$. We denote each feature vector at location (i, j) as a *spatial vector* $\mathbf{f}_{ij} \in \mathbb{R}^d$, which corresponds to a $s \times s$ pixel region in the input image with $s = H/H' = W/W'$.

To serve as a predefined and fixed reference for tamper localization, each anchor vector is sampled from a Rademacher distribution and normalized to have unit norm, constructing $\mathbf{a}_{ij} \in \left\{ -\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}} \right\}^d$, approximating a uniform distribution on the unit hypersphere. Under this construction, the cosine similarity between any clean feature and $\mathbf{a}_{ij}$ is statistically bounded near zero [7, 27], indicating the absence of any intentional alignment. The degree of alignment between $\mathbf{f}_{ij}$ and its corresponding anchor $\mathbf{a}_{ij}$ is then measured by their cosine similarity $s_{ij} = \frac{\mathbf{f}_{ij}^\top \mathbf{a}_{ij}}{\|\mathbf{f}_{ij}\| \|\mathbf{a}_{ij}\|}$.

The perturbation $\boldsymbol{\delta}$ is then optimized to embed the vector-wise localization signal by enforcing $s_{ij} \geq \tau$ across all spatial locations via a hinge loss $\mathcal{L}_{\text{hinge}} = \frac{1}{H'W'} \sum_{i,j} \max(0, \tau - s_{ij})$, which drives each spatial vector to align within the target hypercone margin $\tau \in (0, 1)$. Perceptual losses (*i.e.* $\mathcal{L}_{\text{PSNR}}$ and $\mathcal{L}_{\text{LPIPS}}$) are also applied to minimize perceptual distortion between the clean and perturbed images.

Our objective. Our anchor-aligned perturbation $\boldsymbol{\delta}$ steers image features toward a predefined anchor direction, inducing intentional alignment that serves as a localization signal. This perturbation is designed to be impaired in synthesized foreground regions while remaining stable in background regions. Diffusion models generate foreground content by sampling from a learned data distribution derived from natural images [17]. Consequently, synthesized foreground regions tend to follow this clean distribution and exhibit low cosine similarity with the

anchor. This occurs because the anchor is sampled as a unit vector from the hypersphere, making it statistically unlikely to align with spatial feature vectors extracted from generated images. In contrast, inpainting models reconstruct the background by injecting the original latent representation. This objective preserves the anchor-induced alignment in background regions, even under regeneration. The resulting disparity-aligned background versus misaligned synthesized foreground—creates a measurable contrast in feature space that enables precise spatial localization. We empirically validate these properties in Sec. 5.5.

4.2 Hard Negative Mining Loss and Noisy Perturbation Branch

Hard negative mining loss. In APT, the cosine similarity map is directly used as a localization signal, making spatially uniform alignment critical for reliable detection. However, in general, features from the pretrained image encoder are non-uniformly distributed on the hypersphere [7], causing the initial cosine similarity to vary across spatial locations. We thus propose a hard negative mining loss to further penalize under-optimized spatial vectors. Specifically, we exploit the loss on the top- K spatial vectors with the highest per-vector hinge loss $l_{ij} = \max(0, \tau - s_{ij})$:

$$\mathcal{L}_{\text{hard}} = \frac{1}{|\Omega_K|} \sum_{(i,j) \in \Omega_K} l_{ij} \quad (4)$$

where Ω_K denotes the index set of the top- K spatial vectors sorted by l_{ij} in descending order, with $K = \lceil \rho \cdot H'W' \rceil$ and ρ being the sampling ratio.

Noisy perturbation branch. To enhance robustness of the embedded signal against VAE distortions under background regeneration, we introduce a noisy perturbation branch during optimization. A random binary rectangular mask m_{rand} ($m_{\text{rand}} = 1$ for noise-injected regions) is applied to inject Gaussian noise $\nu \sim \mathcal{N}(0, \sigma^2)$ into the perturbed latent $\tilde{\mathbf{z}}$, yielding a noisy latent:

$$\tilde{\mathbf{z}}_{\text{noisy}} = \tilde{\mathbf{z}} \odot (1 - m_{\text{rand}}) + (\tilde{\mathbf{z}} + \nu) \odot m_{\text{rand}} \quad (5)$$

The noisy latent is decoded as $\tilde{\mathbf{x}}_{\text{noisy}} = \mathcal{D}_{\text{vae}}(\tilde{\mathbf{z}}_{\text{noisy}})$, and both $\mathcal{L}_{\text{hinge}}$ and $\mathcal{L}_{\text{hard}}$ are applied to the resulting noisy image. This strategy effectively improves robustness against latent-space distortions introduced by VAE regeneration.

Final objective function. With the model parameters all frozen, the final objective function for updating the perturbation δ is defined as:

$$\mathcal{L}_{\text{total}} = \sum_{\hat{\mathbf{x}} \in \{\tilde{\mathbf{x}}_{\text{perturb}}, \tilde{\mathbf{x}}_{\text{noisy}}\}} [\mathcal{L}_{\text{hinge}}(\hat{\mathbf{x}}) + \mathcal{L}_{\text{hard}}(\hat{\mathbf{x}})] + \lambda_p \mathcal{L}_{\text{PSNR}} + \lambda_l \mathcal{L}_{\text{LPIPS}}, \quad (6)$$

where $\mathcal{L}_{\text{hinge}}(\hat{\mathbf{x}})$ and $\mathcal{L}_{\text{hard}}(\hat{\mathbf{x}})$ are computed on the cosine similarity map $\hat{s}$ extracted from $\hat{\mathbf{F}}$ via $\mathcal{E}_{\text{img}}$. The full objective jointly enforces spatially uniform anchor alignment across both the perturbed and noisy images, while minimizing perceptual distortion. After iterative gradient updates, we enforce an ℓ_∞ -norm constraint by clipping the pixel-domain perturbation within a budget ϵ_δ . The final perturbed image $\hat{\mathbf{x}}$ is obtained as: $\hat{\mathbf{x}} = \text{Clip}(\mathbf{x} + \text{Clip}(\tilde{\mathbf{x}}_{\text{perturb}} - \mathbf{x}, -\epsilon_\delta, \epsilon_\delta), 0, 1)$.

4.3 Tamper Localization

At verification, given a test image $\tilde{\mathbf{x}}'$, we first extract its feature map via $\mathcal{E}_{\text{img}}$ and compute the per-vector cosine similarity map $s' \in \mathbb{R}^{H' \times W'}$. Average pooling is then applied to s' to smooth per-vector variance, and the resulting map is passed to the detection function $\mathcal{D}$, which produces a pixel-level prediction mask $m' \in \{0, 1\}^{H \times W}$ (i.e. $\mathcal{D}(s') = m'$) via either training-free prediction or a shallow mask decoder, as described below.

Training-free prediction. The detection function $\mathcal{D}$ processes the similarity map to produce a pixel-level prediction mask via bilinear interpolation, temperature scaling, and sigmoid activation $\sigma(\cdot)$:

$$m' = \sigma(T \cdot \text{Upsample}(s')), \quad (7)$$

where T is the temperature parameter. The values 1 and 0 in the predicted mask m' represent perturbed and tampered regions, respectively. A binary localization mask is obtained by thresholding at 0.5.

Shallow mask decoder. To achieve fine-grained localization, we introduce a lightweight shallow mask decoder. Since vector-wise prediction operates independently at each spatial location, it may produce spatially inconsistent masks with isolated false predictions. The decoder addresses this by learning a locality prior that masked regions tend to be spatially contiguous. This allows to leverage surrounding context to suppress isolated errors and fill incomplete predictions.

The decoder $\mathcal{D}_\theta$, with learnable parameters θ , is trained to predict a mask $m' \in \mathbb{R}^{1 \times H \times W}$ from a cosine similarity map $\hat{s}'$, supervised against a ground-truth (GT) mask m_{GT} (1: perturbed, 0: tampered) by a joint loss combining binary cross-entropy (BCE) and Dice loss. Since constructing a large-scale dataset of paired perturbed and tampered images is practically infeasible, we instead generate synthetic similarity maps from clean images alone. Specifically, we construct a synthetic similarity map $\hat{s}'$ to simulate the expected cosine similarity distribution after inpainting, by adding alignment gain to background regions and imposing alignment collapse on foreground regions:

$$\hat{s}' = \underbrace{(s' + \Delta_{\text{bg}})}_{\text{alignment gain}} - \underbrace{\Delta_{\text{fg}} \cdot (1 - m_{\text{GT}})}_{\text{alignment collapse}} + \eta, \quad (8)$$

where $\Delta_{\text{bg}} \sim \mathcal{U}(\alpha_1\tau, \alpha_2\tau)$ models the alignment gain in perturbed regions, and $\Delta_{\text{fg}} \sim \mathcal{U}(\beta_1\tau, \beta_2\tau)$ models the alignment collapse in tampered areas. Here, $\eta \sim \mathcal{N}(0, 0.03^2)$ is additive noise that is introduced to inject per-vector variance in the similarity map. This decoder comprises two upsampling stages ($\times 2$ and $\times 4$), each comprising an upsampling layer followed by two conv-LayerNorm-ReLU blocks.

5 Experiments

5.1 Implementation Details

We optimize an image at resolution $H \times W = 256 \times 256$ with $H' \times W' = 32 \times 32$. We use DINOv3-ConvNeXt-Small [25] as the image encoder with feature layer 1

($d = 192$), and the VAE from `sd-legacy/stable-diffusion-inpainting` [22]. The perturbation is optimized solely on the VAE latent space, without the U-Net. We optimize δ using Adam [13] with $\text{lr} = 1.0$ over 150 steps, with $\lambda_p = 0.1$, $\lambda_l = 0.05$, $\epsilon_\delta = 16/255$, $\tau = 0.1$, $T = 5.0$ and $\rho = 0.1$. The noise scale ϵ for the noisy branch is sampled from $\mathcal{U}(0, 0.25)$. On an NVIDIA RTX 4090 GPU, optimization takes 42.52 seconds per image, averaged over 100 samples. For all experiments, we employ a uniform anchor vector across all spatial locations ($\mathbf{a}_{ij} = \mathbf{a}$ for all i, j) to simplify the optimization process.

Evaluation protocol. We evaluate on the first 100 COCO [16] validation images. Tampered regions are generated by applying inpainting within bounding boxes obtained by expanding object segmentation masks by a factor of 1.2. For fair comparison, we adjust the watermark strength of baselines [24, 28, 32] to match our PSNR. For StableGuard [28], we calculate the PSNR between the watermarked and original images, rather than VAE reconstructions, to ensure evaluation consistency. We report F1, AUC, and IoU as localization evaluation metrics. APT refers to training-free prediction, while APT* employs the shallow decoder for fine-grained mask refinement.

Shallow mask decoder. The optional decoder is trained independently on COCO [16] training images using both segmentation and random rectangular masks as supervision. We train for 6 epochs with batch size 64, $\text{lr} = 3 \times 10^{-4}$, and image size 256×256 . We used $\alpha_1 = 0.3$, $\alpha_2 = 0.9$, $\beta_1 = 0.7$, and $\beta_2 = 1.1$.

5.2 Localization Performance in the FR Setting

Evaluation on SD-Painter. Tab. 1 reports localization performance on Stable Diffusion Inpainting (SD-Painter [22]) in the FR setting. Existing methods such as StableGuard [28] and OmniGuard [32] yield near-random prediction (AUC ≈ 0.5), indicating that their embedded signals are entirely erased by background regeneration. WAM [24] demonstrates modest robustness to FR manipulations, yet its localization remains limited, reaching an IoU of 0.84. In contrast, APT* achieves a superior IoU of 0.92, demonstrating that our vector-wise embedding with latent perturbation preserves localization cues even under the diffusion process. As shown in Fig. 3, APT and APT* consistently yield sharp and spatially accurate predictions, while baselines produce degenerate or uniform masks under FR conditions due to signal loss.

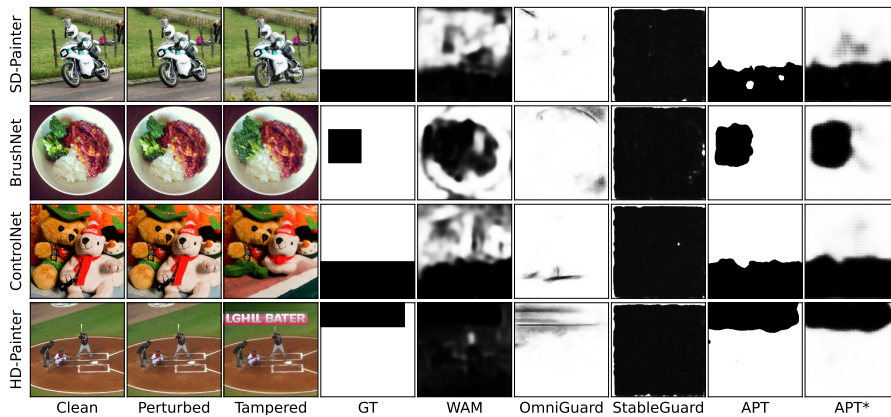
Transferability to diverse inpainting models. We evaluate the generalization of our anchor-aligned perturbation across representative state-of-the-art inpainting models (*e.g.* BrushNet [12], ControlNet [30], and HD-Painter [19]) that utilize diverse background integration strategies detailed in Sec. 2.1. As shown in Tab. 2, APT* maintains a consistent average IoU of 0.93, in line with the results in Tab. 1. This stability stems from leveraging the semi-fragility of aligned vectors rather than visual cues, allowing for consistent localization. While baselines generally follow the failure trends in Tab. 1, WAM [24] shows slightly higher performance on HD-Painter, as HD-Painter tends to generate high-saturation, illustration-like images with distinct boundaries. Overall, these results demon-

Table 1: Quantitative performance in the fully regenerated (FR) setting. **Red:** best, **Blue:** second-best.

	PSNR	F1	AUC	IoU
WAM	32.30	0.91	0.95	0.84
OmniGuard	32.65	0.88	0.49	0.79
StableGuard	25.93	0.12	0.50	0.07
APT	32.80	0.95	0.97	0.90
APT*	32.80	0.96	0.97	0.92

Table 2: Transferability of localization performance in the fully regenerated (FR) setting. **Red:** best, **Blue:** second-best.

	BrushNet		ControlNet		HD-Painter		Average	
	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU
WAM	0.97	0.82	0.96	0.87	0.99	0.87	0.97	0.85
Omni.	0.48	0.79	0.52	0.79	0.56	0.80	0.52	0.79
Stab.	0.50	0.07	0.52	0.07	0.53	0.07	0.52	0.07
APT	0.97	0.89	0.99	0.95	0.97	0.91	0.98	0.91
APT*	0.98	0.92	0.98	0.94	0.97	0.92	0.98	0.93

**Fig. 3:** Qualitative localization comparison under the FR setting across four inpainting models [12, 19, 22, 30]. While baselines produce degenerate or near-uniform predictions due to background signal disruption, both APT and APT* consistently predict accurate masks across diverse inpainting models.

strate that our anchor-aligned perturbation provides a robust and generalizable localization mechanism across modern diffusion-based inpainting frameworks.

5.3 Localization Performance in the SP Setting

Evaluation on SD-Painter. Tab. 3 presents SP localization results on SD-Painter [22]. In the SP setting, all methods achieve near-perfect performance, and both APT and APT* maintain a competitive AUC of 0.99, confirming reliable discrimination between perturbed and tampered regions. The marginal IoU gap relative to baselines stems from the fact that our framework is not explicitly optimized for splicing boundaries, but instead captures generalized tampering cues based on anchor alignment. As shown in Fig. 4, the predicted masks closely follow the ground truth, with discrepancies limited to subtle boundary regions.

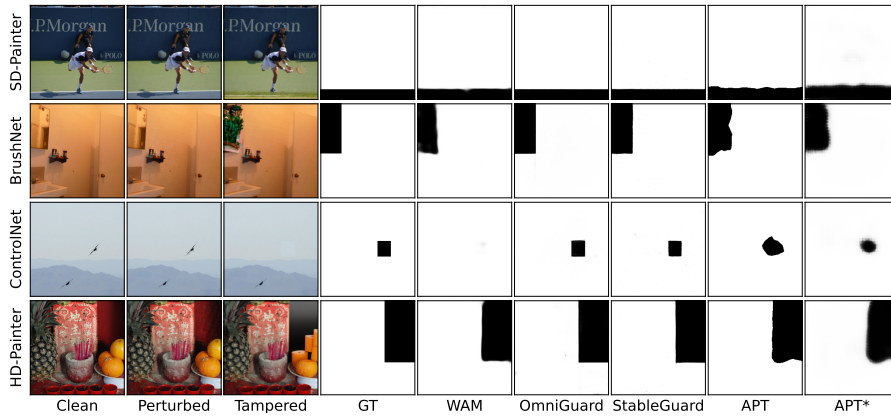
Transferability to diverse inpainting models. We evaluate the generalization across representative inpainting models [12, 19, 30]. As shown in Tab. 4,

Table 3: Quantitative performance in the spliced (SP) setting. **Red**: best, **Blue**: second-best.

	PSNR	F1	AUC	IoU
WAM	32.30	1.00	1.00	0.99
OmniGuard	32.65	1.00	1.00	1.00
StableGuard	25.93	1.00	1.00	1.00
APT	32.80	0.97	0.99	0.94
APT*	32.80	0.97	0.99	0.95

Table 4: Transferability across diverse inpainting models under the spliced (SP) setting. **Red**: best, **Blue**: second-best.

	BrushNet		ControlNet		HD-Painter		Average	
	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU
WAM	1.00	0.99	0.99	0.99	1.00	0.99	1.00	0.99
Omni.	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99
Stab.	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
APT	0.99	0.95	0.99	0.96	0.99	0.96	0.99	0.96
APT*	0.99	0.95	0.99	0.96	0.99	0.94	0.99	0.95

**Fig. 4:** Qualitative localization comparison in the SP setting. All methods produce visually plausible masks, and the APT variants remain highly competitive with state-of-the-art baselines despite not being explicitly optimized for splicing boundaries.

APT and APT* consistently achieve an average AUC of 0.99 across all inpainting architectures, confirming that our approach generalizes well to unseen models. Qualitative comparisons across recent inpainting models are provided in Fig. 4.

5.4 Robustness to Image Corruptions

Tab. 5 evaluates robustness under color jitter (10% brightness increase), Gaussian blur (kernel size 3), and JPEG compression ($Q = 95$) in both FR and SP settings. In the FR setting, APT* maintains strong localization with an average IoU of 0.90, demonstrating consistent robustness against diverse image corruptions. In the SP setting, WAM and OmniGuard benefit from seen corruptions during training (All for WAM; JPEG and color jitter for OmniGuard), yet OmniGuard still collapses under JPEG and blur. While both StableGuard and our framework are evaluated on entirely unseen corruptions, StableGuard collapses to an average AUC of 0.80 in the SP setting, whereas both APT and APT* maintain a robust AUC above 0.96. This demonstrates that the anchor-

Table 5: Robustness evaluation under image corruptions. APT* maintains consistent localization against unseen corruptions, with the shallow decoder effectively recovering signal degradation in challenging cases. **Red**: best, **Blue**: second-best.

	Fully Regenerated (FR)								Spliced (SP)							
	Color Jitter		Gaussian Blur		JPEG		Average		Color Jitter		Gaussian Blur		JPEG		Average	
	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU
WAM	0.94	0.79	0.95	0.86	0.94	0.80	0.94	0.82	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
Omni.	0.48	0.79	0.51	0.79	0.47	0.79	0.49	0.79	0.99	1.00	0.58	0.79	0.47	0.79	0.82	0.93
Stab.	0.50	0.07	0.52	0.07	0.53	0.07	0.52	0.07	0.98	0.94	0.91	0.12	0.52	0.07	0.80	0.37
APT	0.95	0.88	0.95	0.87	0.90	0.64	0.93	0.80	0.98	0.91	0.97	0.92	0.94	0.78	0.96	0.87
APT*	0.96	0.91	0.96	0.91	0.93	0.88	0.95	0.90	0.98	0.93	0.97	0.93	0.95	0.91	0.97	0.93

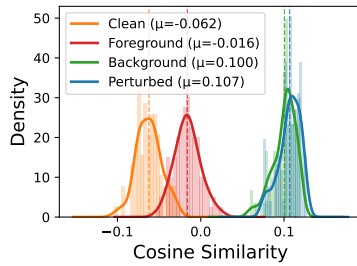


Fig. 5: Cosine similarity distributions empirically validating the alignment-based semi-fragility.

Table 6: Ablation study on each proposed component. $\mathcal{L}_{\text{noise}}$ and $\mathcal{L}_{\text{hard}}$ provide orthogonal gains across both FR and SP settings, with the full model achieving the best localization performance.

$\mathcal{L}_{\text{noise}}$	$\mathcal{L}_{\text{hard}}$	Perceptual	FR			SP		
		PSNR	F1	AUC	IoU	F1	AUC	IoU
		35.71	0.49	0.77	0.35	0.71	0.86	0.57
✓		35.22	0.72	0.86	0.59	0.88	0.93	0.80
	✓	34.77	0.72	0.89	0.61	0.91	0.96	0.86
✓	✓	32.80	0.94	0.97	0.90	0.97	0.99	0.94

alignment serves as a robust signal, with the shallow decoder further recovering localization under challenging corruptions.

5.5 Alignment Disparity

To empirically validate the alignment disparity discussed in Sec. 4.1, we analyze the cosine similarity distributions using 100 COCO validation images. We compare the average similarity scores of perturbed images, clean images, and the foreground and background regions of fully regenerated images tampered with center-mask SD-Painter [22]. As illustrated in Fig. 5, the synthesized foreground regions exhibit a significant deviation from the anchor vectors (mean $\mu=-0.016$), comparable to the low-alignment in clean images ($\mu=-0.062$). In contrast, the background regions of the tampered images maintain a high alignment ($\mu=0.100$) that is nearly identical to that of the perturbed images ($\mu=0.107$). These results demonstrate that foreground regions follow the generative model’s native distribution, exhibiting low anchor alignment.

5.6 Feature Layer Selection

To assess the anchor alignment across the image encoder layers, we optimize perturbations on 100 COCO training images and evaluate cosine similarity maps

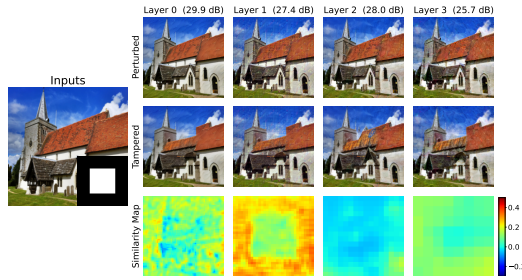


Fig. 6: Cosine similarity maps across image encoder feature layers (0-3) under FR inpainting.

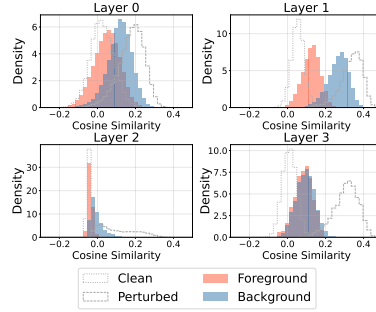


Fig. 7: Cosine similarity distributions across feature layers.

after center-mask SD-Painter FR inpainting. As shown in Fig. 6, Layer 1 exhibits the sharpest spatial boundary, supported by a distinct bimodal separation between foreground (mean $\mu=0.124$) and background ($\mu=0.265$) in Fig. 7. We found that Layer 0 produces high-frequency perturbations that are vulnerable to the VAE bottleneck [22] (see Suppl. for details). In deeper layers (2 and 3), wider receptive fields cause foreground tampering cues to entangle with surrounding background vectors, collapsing the foreground-background disparity. Layer 1 is thus selected for its superior balance of robustness and spatial granularity.

5.7 Ablation Study

Tab. 6 validates the contribution of each proposed component upon the baseline $\mathcal{L}_{\text{hinge}}(\tilde{\mathbf{x}}_{\text{perturb}}) + \lambda_p \mathcal{L}_{\text{PSNR}} + \lambda_l \mathcal{L}_{\text{LPIPS}}$. Here, $\mathcal{L}_{\text{noise}} = \mathcal{L}_{\text{hinge}}(\tilde{\mathbf{x}}_{\text{noisy}})$ denotes the alignment on the noisy perturbation branch, and $\mathcal{L}_{\text{hard}}$ denotes the hard negative mining loss on $\tilde{\mathbf{x}}_{\text{perturb}}$. Each component independently improves localization, with $\mathcal{L}_{\text{noise}}$ improving SP and FR IoU by +0.23 and +0.24, and $\mathcal{L}_{\text{hard}}$ providing a larger boost (+0.29 SP, +0.26 FR). The full model (Eq. (6)) achieves the best performance, demonstrating the independent effectiveness of the two components.

6 Conclusion

We presented APT, a semi-fragile perturbation framework that addresses the fully regenerated (FR) inpainting setting overlooked by existing proactive localization methods. By embedding dense, vector-wise signals via latent perturbation, APT achieves alignment-based semi-fragility—robust to background regeneration yet fragile to foreground synthesis. Experiments across a range of inpainting architectures demonstrate that APT significantly outperforms existing methods in the FR setting while remaining competitive in the SP setting, positioning it as a practical forensic framework generalizable across tampering types unknown at test time.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World), (RS-2023-00237965, Recognition, Action and Interaction Algorithms for Open-world Robot Service), and (No. RS-2020-II200153). Prof. Sung-Eui Yoon and Prof. Soeul Son are co-corresponding authors.

References

1. Asnani, V., Yin, X., Hassner, T., Liu, X.: Malp: Manipulation localization using a proactive scheme. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12343–12352 (2023)
2. Associated Press: Artificial intelligence is being used to spread misinformation about the israel-hamas war (2023), <https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47>, accessed: 2025
3. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. *ACM transactions on graphics (TOG)* **42**(4), 1–11 (2023)
4. Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al.: Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699* (2025)
5. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(3), 3539–3553 (2022)
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
7. Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., Douze, M.: Watermarking images in self-supervised latent spaces. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3054–3058. IEEE (2022)
8. Giakoumoglou, P., Karageorgiou, D., Papadopoulos, S., Petrantonakis, P.C.: Sagi: Semantically aligned and uncertainty guided ai image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16090–16101 (2025)
9. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: 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. pp. 20606–20615 (2023)
10. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3155–3165 (2023)
11. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., Cheng, G.: Sida: Social media image deepfake detection, localization and explanation with

- large multimodal model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28831–28841 (2025)
12. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: European Conference on Computer Vision. pp. 150–168. Springer (2024)
 13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
 14. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. <https://bf1.ai/blog/flux-2> (2025)
 15. Li, D., Zhu, J., Fu, X., Guo, X., Liu, Y., Yang, G., Liu, J., Zha, Z.J.: Noise-assisted prompt learning for image forgery detection and localization. In: European Conference on Computer Vision. pp. 18–36. Springer (2024)
 16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
 17. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.R.: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11461–11471 (2023)
 18. Luo, S., Tan, Y., Patil, S., Gu, D., Von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556 (2023)
 19. Manukyan, H., Sargsyan, A., Atanyan, B., Wang, Z., Navasardyan, S., Shi, H.: Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. In: The Thirteenth International Conference on Learning Representations (2025)
 20. Mareen, H., Karageorgiou, D., Van Wallendael, G., Lambert, P., Papadopoulos, S.: Tgif: Text-guided inpainting forgery dataset. In: 2024 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2024)
 21. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
 22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
 23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
 24. Sander, T., Fernandez, P., Durmus, A.O., Furon, T., Douze, M.: Watermark anything with localized messages. In: The Thirteenth International Conference on Learning Representations (2025)
 25. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
 26. The Guardian: Fake image of pentagon explosion goes viral (May 2023), <https://www.theguardian.com/technology/2023/may/22/pentagon-ai-generated-image-explosion>, accessed: 2025
 27. Vukotić, V., Chappelier, V., Furon, T.: Are deep neural networks good for blind image watermarking? In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–7. IEEE (2018)

28. Yang, H., Liu, B., Xu, X., Xu, C., Yu, Y., Huang, Z., Wang, Y., He, S.: Stableguard: Towards unified copyright protection and tamper localization in latent diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
29. Yu, Z., Ni, J., Lin, Y., Deng, H., Li, B.: Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12765–12774 (2024)
30. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
31. Zhang, X., Li, R., Yu, J., Xu, Y., Li, W., Zhang, J.: Editguard: Versatile image watermarking for tamper localization and copyright protection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11964–11974 (2024)
32. Zhang, X., Tang, Z., Xu, Z., Li, R., Xu, Y., Chen, B., Gao, F., Zhang, J.: Omni-guard: Hybrid manipulation localization via augmented versatile deep image watermarking. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3008–3018 (2025)
33. Zheng, C., Lan, Y., Wang, Y.: Lanpaint: Training-free diffusion inpainting with asymptotically exact and fast conditional sampling. arXiv preprint arXiv:2502.03491 (2025)

Supplementary Material

Table of Contents

A	Additional Results	1
A.1	Evaluation on Segmentation Masks	1
A.2	Evaluation on 500 Images	2
A.3	Transferability across Different VAEs and Architectures	2
B	Additional Analysis	3
B.1	Signal Disruption across Baselines	3
B.2	Layer-wise Perturbation Analysis	4
B.3	Trade-offs	4
C	Implementation Details	5
C.1	Optimization	5
C.2	Shallow Decoder	5
C.3	Evaluation	5
D	Discussion	7
D.1	Threat Model	7
D.2	Limitation	7
D.3	Societal Impact	7

In this supplementary material, we provide additional results, analyses, implementation details, and discussions. Sec. A presents additional evaluations on segmentation masks, 500 inpainting images, and transferability across different VAEs and architectures. Sec. B analyzes baseline signal disruptions and perturbations across encoder layers, and discusses efficiency-quality trade-offs. Sec. C details the optimization algorithm, mask generation for the shallow decoder, and evaluation protocols. Finally, Sec. D discusses our threat model, limitations, and societal impacts.

Throughout this supplementary, we use APT to refer to our method with training-free prediction, and APT* to denote the variant that employs the trained shallow mask decoder. The code will be made publicly available.

A Additional Results

A.1 Evaluation on Segmentation Masks

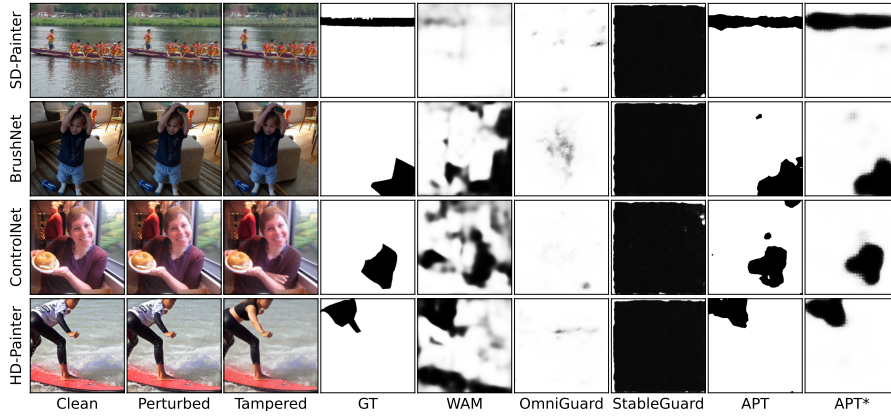
We further evaluate localization performance on images manipulated using instance segmentation masks. Tab. 7 (for SD-Painter) and Tab. 8 (across four inpainting models) show that both APT and APT* consistently outperform all baselines in the challenging fully regenerated (FR) setting. Qualitative results in Fig. 8 further demonstrate the superior localization results of our framework.

Table 7: Quantitative performance in the FR setting on segmentation masks. **Red:** best, **Blue:** second-best.

	PSNR	F1	AUC	IoU
WAM	32.30	0.93	0.96	0.88
OmniGuard	<u>32.65</u>	0.95	0.53	<u>0.91</u>
StableGuard	25.93	0.12	0.46	0.06
APT	32.80	<u>0.96</u>	0.96	0.92
APT*	32.80	0.97	<u>0.95</u>	0.94

Table 8: Transferability of localization performance in the FR setting on segmentation masks. **Red:** best, **Blue:** second-best.

	BrushNet		ControlNet		HD-Painter		Average	
	AUC	IoU	AUC	IoU	AUC	IoU	AUC	IoU
WAM	<u>0.96</u>	0.87	0.95	0.89	0.98	0.88	<u>0.96</u>	0.88
Omni.	0.58	<u>0.91</u>	0.52	0.90	0.55	0.90	0.55	0.90
Stab.	0.47	0.06	0.50	0.06	0.51	0.06	0.49	0.06
APT	0.97	<u>0.91</u>	0.98	0.96	<u>0.95</u>	<u>0.92</u>	<u>0.96</u>	<u>0.93</u>
APT*	0.97	0.94	0.98	0.96	<u>0.95</u>	0.94	0.97	0.95

**Fig. 8:** Qualitative localization comparison under the FR setting on segmentation masks across four inpainting models.

A.2 Evaluation on 500 Images

We scale our evaluation to 500 inpainting samples across both FR and SP settings, utilizing the first 500 COCO [16] validation dataset following prior work [31, 32]. Tab. 9 shows that APT and APT* maintain consistent localization, confirming that our localization performance generalizes to larger test sets.

A.3 Transferability across Different VAEs and Architectures

We further evaluate the transferability of APT* when the protection perturbation is optimized with a specific SD-VAE but applied to inpainting models with different VAEs or generation architectures. To this end, we test APT* on five additional inpainters covering diverse model families: a LoRA-fine-tuned SD model (LCM-LoRA-SDv1-5 [18]), SDXL-inpainting [21], a flow-matching inpainting model (SDv3 [6]), and recent flow-matching T2I models used for inpainting through LanPaint [33], including FLUX.2-klein-9B [14] and Z-Image-Turbo [4]. Although the perturbation δ is optimized for a particular SD-VAE, Tab. 10 shows that APT* transfers effectively across non-identical VAEs and

Table 9: Localization performance on 500 images under both FR and SP settings. **Red:** best, **Blue:** second-best.

	Perceptual Fully Regenerated (FR)				Spliced (SP)		
	PSNR ($\uparrow$)	F1 ($\uparrow$)	AUC ($\uparrow$)	IoU ($\uparrow$)	F1 ($\uparrow$)	AUC ($\uparrow$)	IoU ($\uparrow$)
WAM	32.64	0.89	0.95	0.83	1.00	1.00	<u>0.99</u>
OmniGuard	<u>33.17</u>	0.86	0.50	0.77	<u>0.99</u>	1.00	<u>0.99</u>
StableGuard	26.61	0.12	0.50	0.07	1.00	1.00	1.00
APT	33.44	<u>0.93</u>	<u>0.97</u>	<u>0.89</u>	0.97	<u>0.99</u>	0.94
APT*	33.44	0.95	0.98	0.91	0.97	<u>0.99</u>	0.95

Table 10: Transferability performance in the fully regenerated (FR) setting. Results are reported as AUC/IoU. We use the same SD-VAE-optimized perturbation δ and the same shallow decoder for APT*.

Inpainting Method	VAE Type ($H=W, C$)	APT ($\uparrow$)	APT* ($\uparrow$)
<i>Fine-tuned SD</i>			
LCM-LoRA ('23)	SD-VAE (64, 4)	0.99/0.94	0.98/0.94
<i>Non-identical VAEs</i>			
SDXL-inpaint ('23)	SDXL-VAE (128, 4)	0.99/0.94	0.98/0.95
<i>Flow-matching models</i>			
SDv3-inpaint ('24)	SD3-AE (128, 16)	0.97/0.88	0.97/0.91
FLUX.2 ('25)	FLUX.2-AE (128, 32)	0.97/0.92	0.97/0.94
Z-Image-Turbo ('25)	FLUX-AE (128, 16)	0.99/0.95	0.98/0.96

recent inpainting architectures, achieving $\text{AUC} \geq 0.97$ and $\text{IoU} \geq 0.91$. These results indicate that APT* is not limited to the protection-time SD-VAE and can generalize to non-identical VAEs and deterministic ODE-based generation pipelines.

B Additional Analysis

B.1 Signal Disruption across Baselines

While Fig. 1 in the main paper visualizes the embedding and background signals for WAM [24] and APT*, we extend this comparison to other baselines here. Fig. 9 shows that methods like WAM [24] and OmniGuard [32], which embed their signals in the pixel and frequency domains, undergo severe alterations in their localization cues in the FR setting. Although both StableGuard [28] and APT* maintain their signals across both SP and FR settings by leveraging latent space embedding, StableGuard struggles with precise localization due to its reliance on a global embedding, whereas our dense vector-wise approach ensures accurate spatial boundaries.

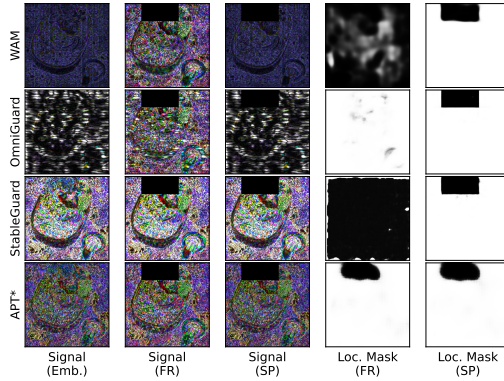


Fig. 9: Extended comparison of embedded signals and localization masks across all baselines under FR and SP settings (cf. Fig. 1).

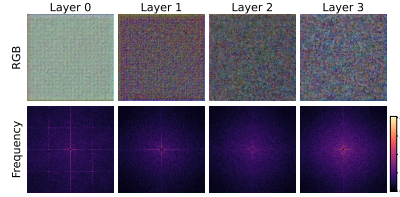


Fig. 10: Average perturbations visualized in RGB and frequency domains across encoder layers 0-3.

B.2 Layer-wise Perturbation Analysis

Expanding on the discussion in Sec. 5.6, we further analyze the perturbations across different image encoder layers. Specifically, we optimize perturbations using only the alignment loss (without the hinge) and perceptual losses to maximize the perturbation strength. The average perturbations in both the RGB and frequency domains are visualized in Fig. 10. Layer 0 exhibits fine-grained horizontal and vertical patterns, which manifest as distinct crosses in the high-frequency spectrum. Layer 1 also forms a cross, but its patterns are coarser and predominantly low-frequency. Layers 2 and 3 concentrate their energy almost exclusively in the low-frequency spectrum, lacking distinct edge patterns in the RGB domain. Layer 0 perturbations are dominated by high frequencies that struggle to propagate the VAE bottleneck [22], while the wider receptive fields of Layers 2–3 hinder precise boundary localization. Layer 1 strikes the right balance, producing coarse structural patterns that survive the VAE and retain sufficient spatial resolution.

B.3 Trade-offs

We investigate the trade-offs among computational efficiency, visual quality (PSNR), and localization performance (see Fig. 11). First, varying the target cosine similarity τ shifts the PSNR-IoU balance predictably (left). Second, tightening the perturbation budget ϵ_δ improves perceptual quality at a modest cost to localization accuracy (middle). Finally, reducing the optimization steps to 50 accelerates the process to 14.34 seconds per image while maintaining highly competitive IoU scores in both settings (right). These results suggest that APT* flexibly adapts to diverse computational and perceptual constraints without substantial performance loss.

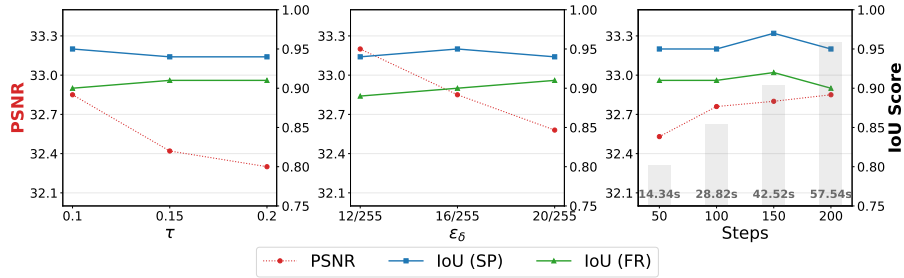


Fig. 11: Trade-off analysis varying target cosine similarity τ (left), perturbation budget ϵ_δ (middle), and optimization steps (right).

C Implementation Details

C.1 Optimization

We detail the complete procedure for our anchor-aligned perturbation optimization in Algorithm 1.

C.2 Shallow Decoder

Mask Generation. During the training of the shallow mask decoder on the COCO dataset, we employ two mask generation strategies with equal probability: (i) unpaired masks and (ii) random rectangular masks. For unpaired masks, we randomly sample an instance segmentation mask from a different image in the dataset. For random rectangular masks, we generate a bounding box with its width and height uniformly sampled from 10% to 50% of the image dimensions. Additionally, we invert the generated mask with a 50% probability. Then, we define the ground-truth mask as $m_{GT} = 1 - \text{mask}$ to represent perturbed regions as 1 and tampered regions as 0.

C.3 Evaluation

Metrics. We evaluate pixel-level localization performance using the F1 score, Area Under the Curve (AUC), and Intersection over Union (IoU). F1 and IoU measure the absolute overlap between the predicted and ground-truth masks at a fixed threshold (0.5). In contrast, AUC is a threshold-independent metric that evaluates the model’s inherent ability to separate tampered and perturbed pixels.

This distinction is crucial due to the class imbalance in manipulated images, where the tampered region is often much smaller than the intact background. For instance, a collapsed model that blindly predicts the entire image as the majority class (perturbed) will yield a near-random AUC (≈ 0.5) but may produce misleadingly high F1 and IoU scores. Therefore, AUC is essential to verify the model’s true discriminative power.

Algorithm 1 APT Localization-Mark Embedding Optimization

Input: Original image x , Pretrained VAE Encoder $\mathcal{E}_{vae}$, VAE Decoder $\mathcal{D}_{vae}$, Image Encoder $\mathcal{E}_{img}$, Perturbation budget ϵ_δ , Optimization steps $Steps$

Output: Protected image $\tilde{x}_{perturb}$

```

1:  $z \leftarrow \mathcal{E}_{vae}(x)$ 
2: Initialize learnable latent perturbation  $\delta \leftarrow \mathbf{0}$ 
3: for  $t = 1$  to  $Steps$  do
4:   Sample random binary rectangular mask  $m_{rand}$ 
5:   Sample noise scale  $\sigma \sim \mathcal{U}(0, 0.25)$ 
6:   Sample Gaussian noise  $\nu \sim \mathcal{N}(0, \sigma^2)$ 
7:    $\tilde{z} \leftarrow z + \delta$ 
8:    $\tilde{z}_{noisy} \leftarrow \tilde{z} \odot (1 - m_{rand}) + (\tilde{z} + \nu) \odot m_{rand}$ 
9:    $\tilde{x}_{perturb} \leftarrow \mathcal{D}_{vae}(\tilde{z})$ 
10:   $\tilde{x}_{noisy} \leftarrow \mathcal{D}_{vae}(\tilde{z}_{noisy})$ 
11:  Compute  $\mathcal{L}_{hinge}(\hat{x})$  and  $\mathcal{L}_{hard}(\hat{x})$  for  $\hat{x} \in \{\tilde{x}_{perturb}, \tilde{x}_{noisy}\}$ 
12:  Compute perceptual losses  $\mathcal{L}_{PSNR}(\tilde{x}_{perturb}, x)$  and  $\mathcal{L}_{LPIPS}(\tilde{x}_{perturb}, x)$ 
13:   $\mathcal{L}_{total} \leftarrow \sum_{\hat{x} \in \{\tilde{x}_{perturb}, \tilde{x}_{noisy}\}} [\mathcal{L}_{hinge}(\hat{x}) + \mathcal{L}_{hard}(\hat{x})] + \lambda_p \mathcal{L}_{PSNR} + \lambda_l \mathcal{L}_{LPIPS}$ 
14:  Update  $\delta$  by minimizing  $\mathcal{L}_{total}$ 
15: end for
16:  $\tilde{x} \leftarrow \mathcal{D}_{vae}(z + \delta)$ 
17:  $\tilde{x}_{perturb} \leftarrow \text{Clip}(x + \tilde{x} - x, -\epsilon_\delta, \epsilon_\delta), 0, 1)$ 
18: return  $\tilde{x}_{perturb}$ 

```

While AUC indicates discriminative power, a high IoU ensures practical usability. In real-world forensic analysis, verifiers rarely know in advance whether an image has been spliced or fully regenerated. A model might achieve high AUC in both settings by separating the classes well internally, but if its output score distributions shift drastically between SP and FR, a single fixed threshold will yield suboptimal results. Thus, a robust forensic framework must deliver high IoU at a constant, predefined threshold across both tampering types. APT and APT* maintain consistent output probability distributions across both settings (see Tab. 1 and Tab. 3), allowing a single threshold to generalize effectively.

Other baselines. To ensure a fair comparison, we adjust the embedding strengths of the baseline methods to align their perceptual quality with ours. Specifically, we set the watermark strength to 3.0 and 2.0 for WAM [24] and OmniGuard [32], respectively. Furthermore, StableGuard [28] originally reports PSNR between the VAE-reconstructed image and the watermarked image, whereas we strictly compute the PSNR directly between the original clean image and the protected image across all methods. This ensures a consistent and practical evaluation of perceptual distortion from embedding signals.

Inpainting models. All evaluated models [12, 19, 22, 30] synthesize images using 50 inference steps and an empty text prompt to fill the masked regions to blend naturally with the surrounding context.

D Discussion

D.1 Threat Model

We consider three distinct entities: the *image owner* (defender), the *attacker*, and the *verifier*. The image owner protects the original image by injecting an imperceptible perturbation prior to public distribution. To model a practical defense scenario, we assume the detector operates as a black-box API. While the detection algorithm is publicly accessible for integrity verification, the internal anchor vectors remain securely hidden on the server side as a secret key. This separation prevents attackers from trivially extracting the anchors or bypassing the detection mechanism.

D.2 Limitation

APT relies on the alignment disparity between the protected background and the synthesized foreground. If an adversary performs a copy-move forgery within the same protected image, the copied region retains the embedded anchor alignment. This lack of alignment disparity can potentially lead to false negatives.

D.3 Societal Impact

APT expands the scope of proactive forensics by addressing the fully regenerated setting, a challenging yet realistic scenario in state-of-the-art localized image synthesis. By providing a reliable baseline for this setting, we contribute to the broader effort to ensure practical and robust image authentication in the wild.

However, a potential negative impact exists if the framework is misused. A malicious actor could apply the APT perturbation to an already-manipulated image, embedding a valid localization signal that misleads the verification system into authenticating the forgery as an intact original. To mitigate this risk, future research should explore integrating proactive defenses directly at the point of capture (*e.g.*, camera hardware).