Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Scene-Guard preserves scene-manipulation disruption under JPEG recompression

Min Shi

wu9896893690163.com

Chinese People's Armed Police Force Engineering University

Yingnan Zhang

Chinese People's Armed Police Force Engineering University

Hemin Yin

Chinese People's Armed Police Force Engineering University

Yanzhe Zhang

Chinese People's Armed Police Force Engineering University

Article

Keywords: image forensics, generative image editing, compression-aware adversarial protection, scene inpainting, JPEG recompression, proactive defense

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10765213/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Scene-Guard preserves scene-manipulation disruption under JPEG recompression

Min Shi[1,2*], Yingnan Zhang[1,2], Hemin Yin[1,2] and Yanzhe Zhang[1,2]

[1] Engineering University of the Chinese People's Armed Police Force, Xi'an, Shaanxi 710000, China.
[2] Key Laboratory, College of Cryptography Engineering, Engineering University of PAP, Xi'an, Shaanxi 710000, China.

*Corresponding author(s). E-mail(s): wu989689369@163.com;

Abstract

Generative image editors can alter scene content while remaining visually plausible, motivating proactive methods that shift downstream manipulation outputs away from clean-input behavior. We define protection operationally as measurable protected-versus-clean output divergence, not guaranteed edit failure. Scene-Guard centers on the Scene Structural Disruption Module (SSDM), which learns bounded pre-release perturbations for scene inpainting and completion. On the full 12,614-image CASIA v2 source corpus, using four established editors (LaMa, DeepFillv2, Global–Local and MAT) plus one untrained internal ResNet50 architectural surrogate under a fixed centered-square mask, the primary proof-of-concept SSDM point reaches mean SDR 0.500 and retains 88.9–95.4% after JPEG recompression at qualities 50–90. Fidelity is 23.20 dB PSNR and 0.466 SSIM, so this is a disruption–fidelity proof of concept rather than an imperceptible deployment point. On a common 2,000-image reference protocol, one-pass Scene-Guard reaches SDR 0.430 (across-target SD 0.013). In a separate paired generic-perturbation control, a sign perturbation approaches Scene-Guard's raw SDR (0.431 vs 0.436) but not SemDR (0.058 vs 0.794), showing that strong generic distortion can reproduce raw structural divergence without the same semantic shift. Leave-one-target-out folds show no systematic held-out degradation within the controlled panel. CRPS is analyzed separately as a compression-aware training extension, and JAFM as a non-source-disjoint auxiliary double-JPEG protocol study. The evidence therefore supports persistent cross-target output divergence under the fixed-mask setting, but not mask-geometry generalization, universal edit prevention, imperceptible deployment or robustness to arbitrary external editors.

Keywords: image forensics, generative image editing, compression-aware adversarial protection, scene inpainting, JPEG recompression, proactive defense

1 Introduction

Generative image models, including diffusion, GAN and transformer-based editors, enable plausible manipulation of objects, backgrounds and contextual evidence. Here, scene-level manipulation denotes semantic inpainting or completion that changes scene content without relying on identity cues, which is relevant when the scene itself carries evidentiary value.

Safeguards are broadly passive or proactive. Passive methods inspect distributed images for editing or re-encoding traces, including double-JPEG artifacts in DCT statistics [1]. Proactive methods modify images before release so downstream outputs differ from those produced from clean inputs. We use protection in this operational sense—measurable manipulation-output divergence rather than guaranteed edit failure. Prior perturbation defenses address face, diffusion and instruction-guided editing [2–4], but routine recompression is often underexplored even though JPEG can attenuate adversarial perturbations [5].

Scene-Guard is a proactive scene-protection framework centered on SSDM. SSDM optimizes semantic, geometric, scene-level and multi-scale discrepancies over four established inpainting/completion editors plus one untrained internal architectural surrogate; SDR quantifies the resulting protected-versus-clean output difference. The surrogate is an auxiliary target, not a representative editor. CRPS is a separate compression-aware training extension, and JAFM a separate auxiliary single-versus-double JPEG branch; neither is used to support the main SSDM performance claim or a coupled end-to-end system.

The evaluation follows this hierarchy. The primary SSDM checkpoint is used for output divergence, JPEG survivability, target-wise consistency and a unified-protocol reference comparison. It is interpreted as a proof-of-concept point on a disruption–fidelity trade-off. Five leave-one-target-out folds vary target architecture while holding a fixed centered-square mask constant; realistic-editor conclusions rely primarily on the four established editors, with the untrained surrogate treated only as an auxiliary target. CRPS is studied by matched retraining and JAFM by a separate full-corpus forensic protocol, so results from distinct regimes are not pooled into claims of universal protection or a validated coupled system. Supplementary Fig. S1 summarizes the protocol.

2 Results

2.1 Cross-target output divergence persists after JPEG recompression

The primary SSDM point was evaluated on 12,614 CASIA v2 images with LaMa, DeepFillv2, Global–Local, MAT and an internal ResNet50 encoder-decoder whose decoder is randomly initialized and not task-trained; the latter is only an auxiliary architectural target. All targets use the same fixed centered-square mask. Mean SDR over the five-target panel is 0.500, with retained disruption of 0.889, 0.944 and 0.954 after JPEG quality 50, 75 and 90, respectively. Fidelity is 23.20 dB PSNR and 0.466 SSIM, and target-specific SDR spans 0.473–0.518. SDR measures protected-versus-clean manipulation-output divergence, not task failure. CASIA authentic/tampered labels are unused: all 12,614 images are treated as source scenes under the same protocol, so the aggregate is full-corpus rather than authentic-only. Mask geometry is not varied. Target-wise and JPEG results are in Supplementary Table S1.

On the common 2,000-image reference run, one-pass Scene-Guard reaches mean SDR 0.430 with across-model SD 0.013 (Table 1). This is a protocol-specific reference, not a state-of-the-art ranking. In a separate matched refinement run, ten test-time PGD steps raise SDR from 0.441 to 0.456 while PSNR decreases from 22.38 to 22.31 dB; the 0.456 result is therefore not compared directly with Table 1. These two operating points illustrate the disruption–fidelity trade-off. Target-wise and paired refinement results are in Supplementary Tables S2 and S3.

Table 1 Reference comparison on the common 2,000-image unified-adaptation run. The five-target panel contains four established editors and one auxiliary untrained ResNet50 surrogate. Mean/worst SDR use target-specific means; SD is across-target dispersion, not run-to-run uncertainty. The table is a protocol reference, not a state-of-the-art ranking.

Method	Mean SDR	Worst SDR	SD	PSNR (dB)
No defense	0.000	0.000	0.000	–
CMUA-Watermark	0.434	0.404	0.025	23.62
PhotoGuard	0.385	0.361	0.021	23.88
IAP	0.397	0.372	0.024	23.81
ID-Guard (adapted)	0.440	0.407	0.022	23.54
Scene-Guard (G)	0.430	0.408	0.013	22.38

Sample-level visualizations complement the aggregate metrics by showing perturbation distribution and downstream differences around edited regions; they illustrate output-divergence patterns rather than task success or localization (Fig. 1; additional examples in Supplementary Fig. S2).

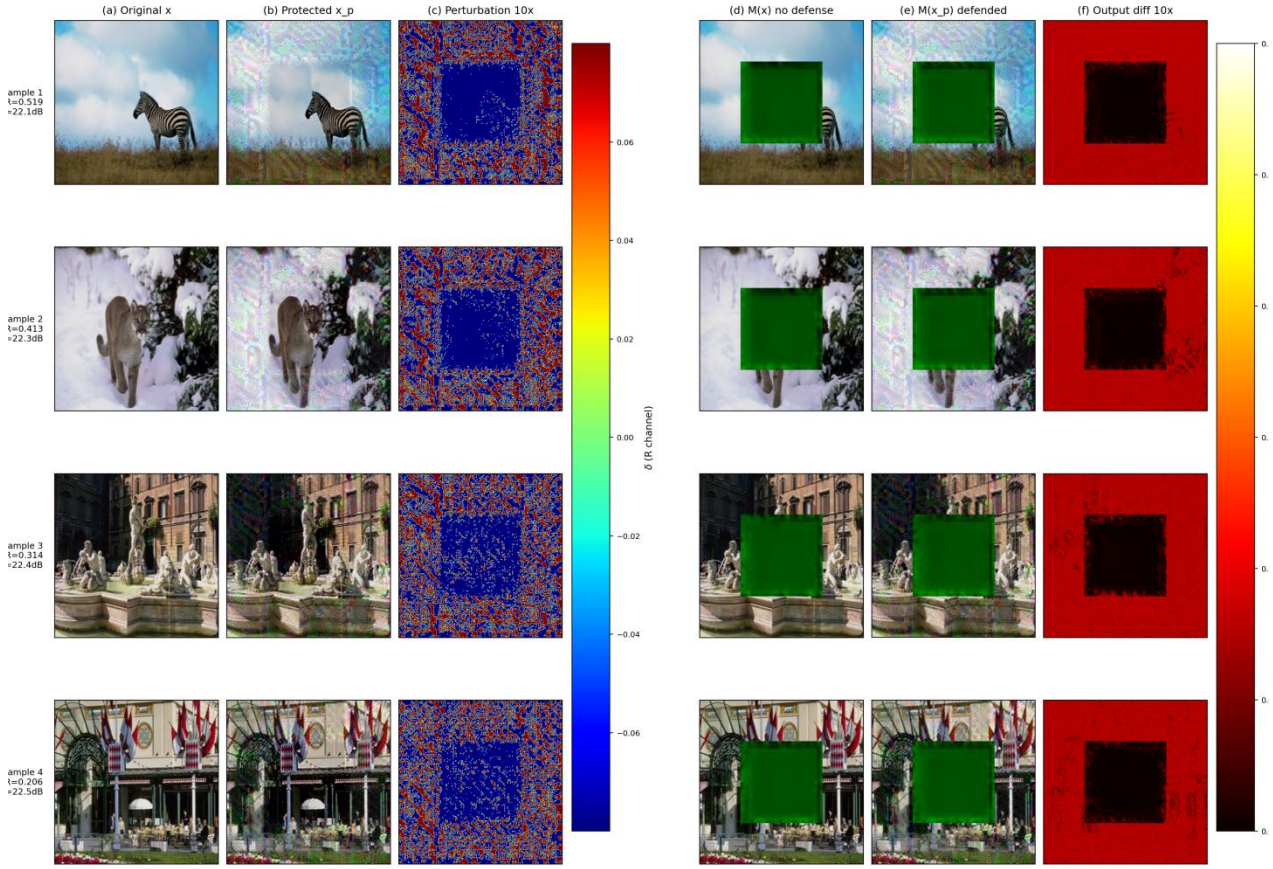


Fig. 1 Representative output-divergence examples. Original/protected inputs, 10× perturbations, unprotected/protected manipulation outputs and 10× output differences are shown under the same fixed centered-square mask. The examples illustrate perturbation and output-difference patterns, not task-success, localization or mask-generalization benchmarks. Source images are from the CASIA v2 authentic subset [11].

2.2 Target-withholding transfer under the fixed-mask protocol

We retrained five leave-one-target-out folds, withholding one target from the generator objective and using it only at test time. The panel contains four established editors and one auxiliary untrained ResNet50 surrogate, and the fixed centered-square mask is unchanged; thus the experiment varies target architecture while holding editing geometry constant. Held-out SDR is at least 99.0% of the mean on the four seen targets in every fold, with means 0.4813 (held out) and 0.4682 (seen). Ratios above 100% reflect target difficulty, not improved generalization. The study shows no systematic degradation within this controlled panel and does not establish transfer to arbitrary external editors or different masks.

Table 2 Leave-one-target-out folds within the fixed-mask panel. Relative SDR is held-out SDR divided by the mean on the four seen targets; values above 100% reflect target difficulty, not improved generalization. The ResNet50 target is an auxiliary untrained surrogate.

Held-out model	Unseen SDR	Seen-target SDR	Relative SDR
LaMa	0.4664	0.4420	105.5%
DeepFillv2	0.4883	0.4932	99.0%
Global–Local	0.4703	0.4598	102.3%
MAT	0.5095	0.4760	107.1%
ResNet50 surrogate	0.4722	0.4700	100.5%
Average	0.4813	0.4682	102.8%

2.3 Matched CRPS retraining isolates JPEG-exposure effects

Because CRPS is absent from the primary SSDM checkpoint, we use five matched 15-epoch retraining variants to study a separate mechanism question: SSDM only, JPEG exposure, JPEG+consistency, JPEG+frozen-JAFM regularization and full CRPS. Without JPEG-aware training, post-JPEG SDR falls from 0.4699 at quality 90 to 0.4479 at quality 50. JPEG exposure alone raises quality-50 SDR to 0.4701 and contracts the quality-50–90 range from 0.0220 to 0.0032 (~86%). The consistency variant reaches the highest quality-50 SDR (0.4704); JAFM regularization slightly improves SSIM while modestly lowering JPEG-averaged SDR; full CRPS gives the highest mean per-sample compression-survival score (CS=1.015). These results characterize cross-quality stabilization, not the origin of the primary checkpoint's JPEG survivability.

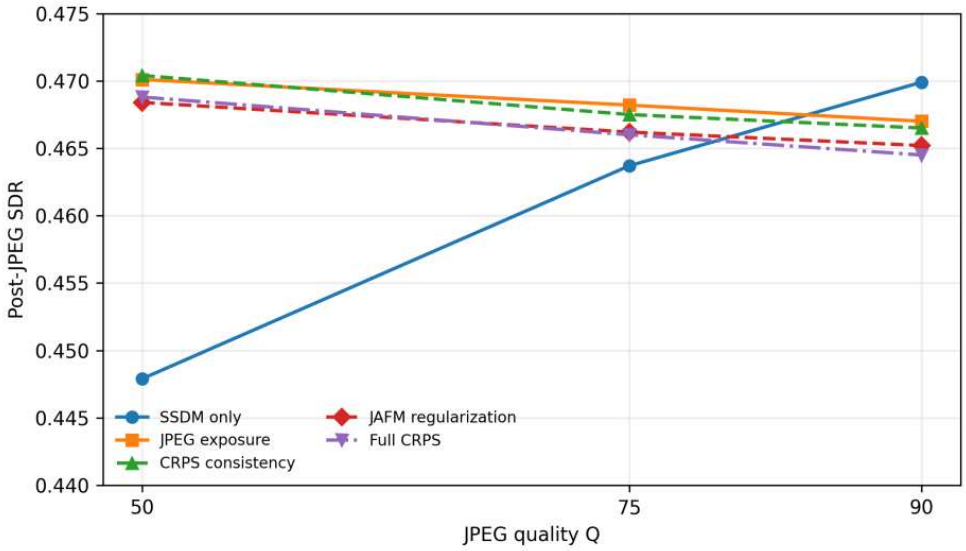


Fig. 2 Matched compression-aware retraining. Post-JPEG SDR for five variants at qualities 50, 75 and 90 shows that explicit JPEG exposure removes most quality-dependent decline in this controlled study; auxiliary losses mainly adjust the stability–fidelity balance. This is a CRPS mechanism study, not an explanation of the primary checkpoint.

Target responses differ: JPEG-aware variants improve DeepFillv2, Global–Local, MAT and the ResNet50 surrogate relative to SSDM-only, whereas LaMa remains strongest with SSDM-only. The claim is therefore limited to improved cross-quality SDR stability in matched CRPS retraining, not uniform gains on every target.

2.4 Auxiliary JAFM protocol characterization

JAFM is reported separately from the main proactive results. Under five fixed quality-factor pairs, 12,614 source images yield 126,140 single/double-JPEG samples. Mean accuracy is 88.28%, ranging from 96.57% (50/85) to 77.42% (85/95), with precision ≥0.912 (Table 3). Because the full-corpus protocol reuses source images represented during

JAFM development, these values characterize the protocol rather than source-disjoint generalization. Active-passive compatibility is also untested.

Table 3 JAFM full-corpus double-JPEG protocol characterization. Each quality pair contains 25,228 samples (12,614 sources × single/double compression); values are not source-disjoint generalization estimates.

QF pair	Accuracy (%)	Precision	Recall	F1
50/75	93.94	0.912	0.973	0.941
50/85	96.57	0.949	0.984	0.966
75/85	89.29	0.941	0.838	0.887
75/95	84.18	0.927	0.742	0.824
85/95	77.42	0.912	0.607	0.729
Mean	88.28	–	–	–

2.5 Disruption–fidelity trade-off

Across a separate 1,000-image sweep (ε=0.03, 0.05, 0.08, 0.12, 0.16; 10 PGD steps for G+PGD), larger perturbation budgets raise SDR but lower PSNR (Fig. 3). The reported ε=0.08 setting is a proof-of-concept point, not a deployment optimum. At this budget refinement increases structural divergence with a small paired PSNR change. Component ablation also shows why SDR must be read with SemDR and fidelity: removing the semantic adversarial term raises SDR from 0.2862 to 0.3968 while SemDR collapses from 0.2627 to 0.0070 and PSNR falls by almost 3 dB.

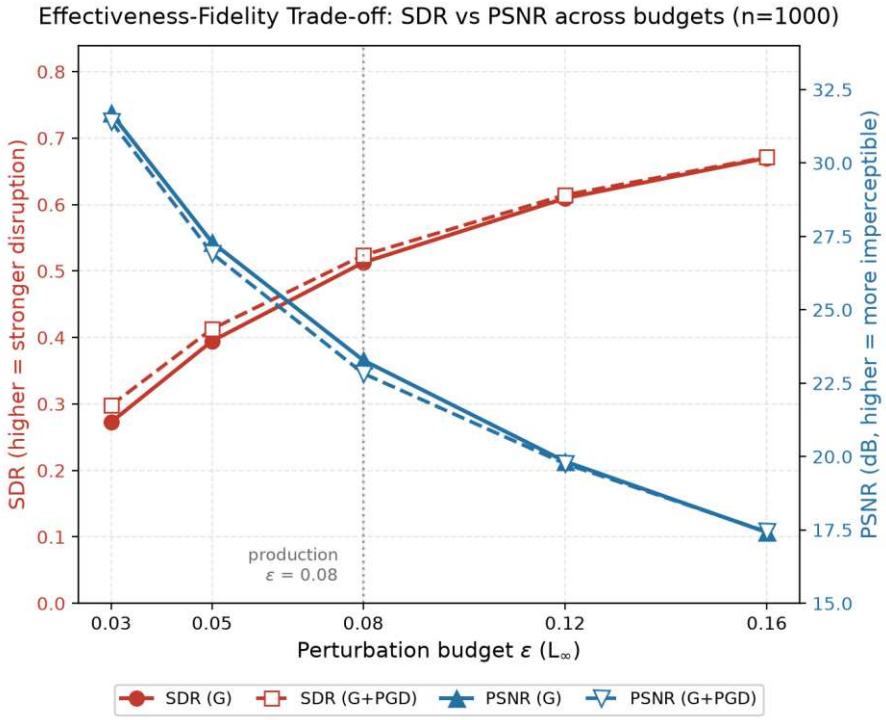


Fig. 3 Disruption–fidelity trade-off (n=1,000). SDR rises and PSNR falls as ε increases; G+PGD uses 10 steps. The dotted line marks the proof-of-concept ε=0.08 setting. Points are raw averages from this separate sweep and do not define a deployment-optimal threshold.

Supplementary Tables S4–S7 provide component, compression-aware and generic-perturbation controls. In a separate paired 2,000-image run at ε=0.08 (Supplementary Table S7), uniform and Gaussian-matched controls produce lower SDR (0.276 and 0.330) with very small SemDR (0.030 and 0.039). A sign control approaches Scene-Guard's raw SDR (0.431 vs 0.436) but has SemDR 0.058 versus 0.794, while source fidelity is slightly worse (22.16 dB/0.466 vs 22.38 dB/0.543). Thus strong generic distortion can reproduce much of raw SDR without the semantic output shift observed for SSDM, reinforcing joint interpretation of SDR, SemDR and fidelity. Separately, the component ablation shows that removing the semantic term raises raw SDR while collapsing SemDR, and matched CRPS retraining shows that JPEG exposure drives most cross-quality stabilization.

3 Discussion

Scene-Guard is therefore interpreted as a proof-of-concept SSDM evaluation, not a closed-set ranking, deployment-ready fidelity claim or combined three-module system. Protection denotes measurable protected-versus-clean manipulation-output divergence. The controlled panel contains four established editors and one untrained internal ResNet50 surrogate used only as an auxiliary architectural target. Across the panel, SSDM produces relatively consistent output divergence that largely persists after JPEG recompression. Leave-one-target-out folds show no

systematic degradation when a target is withheld, but they remain a fixed-mask architecture-withholding study rather than evidence of transfer to arbitrary unseen editors. Separately, matched CRPS retraining shows that differentiable-JPEG exposure reduces quality-dependent SDR variation; because CRPS is absent from primary training, this does not explain the primary checkpoint's JPEG behavior.

The contribution is deliberately scoped to proof-of-concept output divergence across four established inpainting/completion editors under a fixed editing geometry, with the untrained surrogate providing only an auxiliary architecture probe. It does not establish universal protection against generative editing. Table 1 is likewise a unified-protocol reference rather than a state-of-the-art claim (Supplementary Fig. S3 gives a conservative scope comparison).

The experiments also expose a clear disruption–fidelity trade-off. At ε=0.08, PSNR is 23.20 dB and SSIM 0.466, so the primary point is not presented as imperceptible. Test-time PGD increases divergence at additional fidelity cost. Both the semantic-loss ablation and the paired generic-perturbation control show why raw SDR is insufficient: the sign control nearly matches Scene-Guard's SDR (0.431 vs 0.436) but yields SemDR 0.058 rather than 0.794 while providing no fidelity advantage. Generic distortion can therefore drive structural divergence without reproducing the semantic shift observed for SSDM. SDR should be interpreted jointly with SemDR and source fidelity, not as proof that an intended edit failed. JAFM remains an auxiliary non-source-disjoint protocol study and is not evidence of a validated detector-protector pipeline.

Limitations include robustness beyond JPEG (e.g., resizing, cropping, denoising, colour conversion, compound processing and newer codecs), variation in mask shape/size/location, and transfer beyond the four established editors. The five-target panel includes one untrained surrogate that is not equivalent to an established editor. The CASIA aggregate also pools authentic and tampered source images because labels do not enter SSDM or SDR; no authentic-only performance is claimed. JAFM still requires source-disjoint evaluation and direct active-passive testing. Thus, the supported conclusion is narrow: SSDM provides proof-of-concept protected-versus-clean output divergence under a fixed-mask protocol, with substantial JPEG persistence and a measurable fidelity trade-off; CRPS and JAFM remain separate auxiliary studies.

4 Methods

4.1 Threat model and training regimes

Let x be a source image in [0,1] with shape 3×H×W and {M_k} a set of frozen target models. A generator Gθ produces δ=Gθ(x) and x_p=clip(x+δ,0,1), with ||δ||∞≤ε. The objective induces divergence between manipulation outputs from protected and clean inputs; it does not guarantee failure of a specific semantic edit. Gradients are available through active surrogate models and disruption-loss features, giving a white/gray-box setting for that set. In leave-one-target-out folds, the held-out target contributes no training gradient.

Scene-Guard centers on SSDM, which supplies the primary proof-of-concept results using stabilized multi-target weighting. CRPS is a separate compression-aware extension that inserts differentiable JPEG and optionally enables consistency and frozen-JAFM feature terms. JAFM is an independent auxiliary forensic branch and remains frozen when used for CRPS regularization. These regimes are represented by Eq. (1).

$$\min_\theta \ \mathbb{E}_x\left[\mathcal{L}_{\text{SSDM}} + \lambda_c \mathcal{L}_{\text{CRPS}} + \lambda_f \mathcal{L}_{\text{for}}\right], \qquad \| \delta \|_\infty \leq \varepsilon. \tag{1}$$

Setting the compression-specific coefficients in Eq. (1) to zero recovers the primary SSDM regime; they are enabled only in matched CRPS retraining. Fig. 4 summarizes the primary SSDM path.

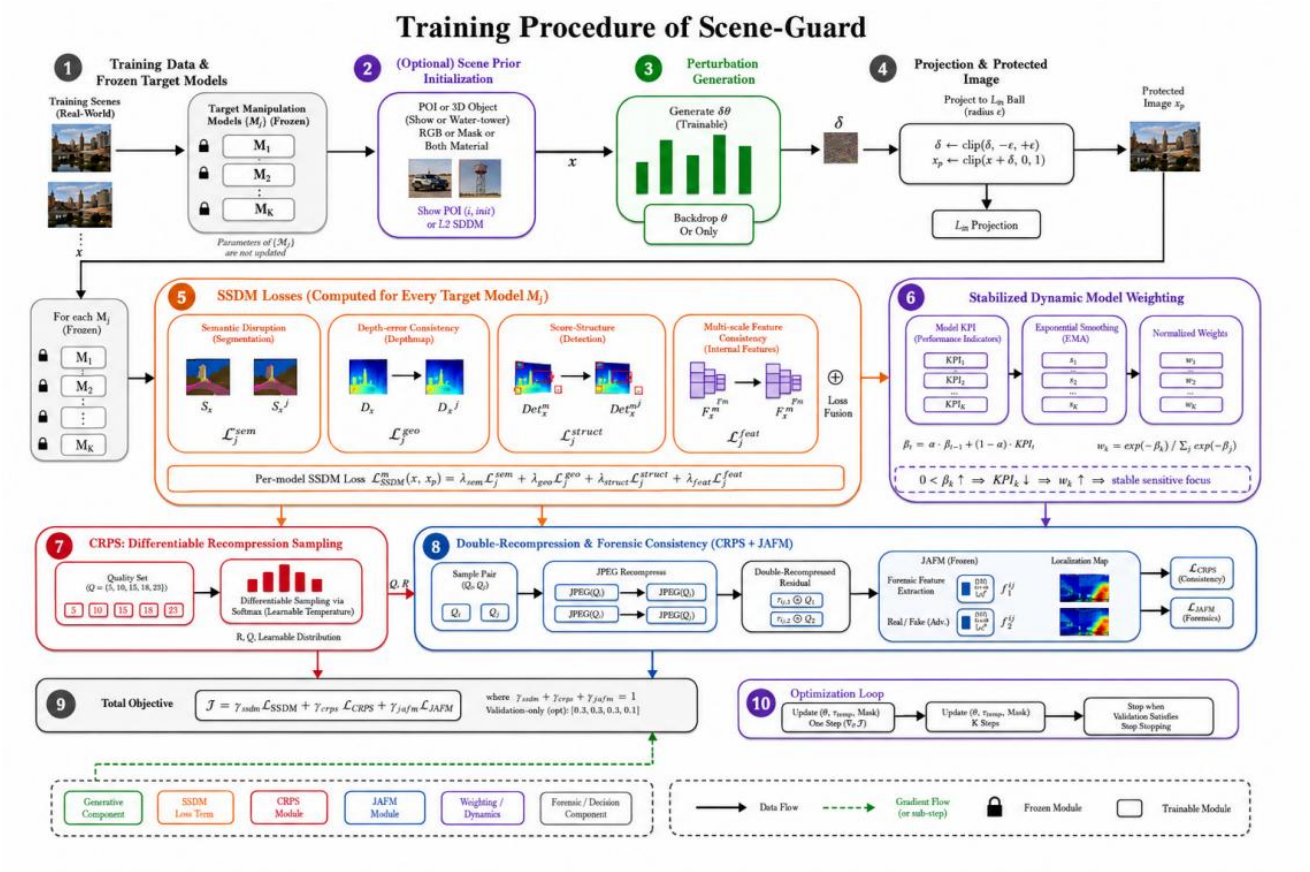


Fig. 4 Primary SSDM training path. The generator produces a bounded perturbation; frozen targets provide semantic, geometric, scene-structure and multi-scale disruption losses combined by stabilized multi-target weighting. CRPS terms are omitted from the primary checkpoint and studied separately; JAFM is an auxiliary branch.

4.2 Scene Structural Disruption Module

SSDM uses Panoptic-FPN/ResNet-50 semantic masks [6], MiDaS depth priors [7] and Places365 scene representations [8]. For each frozen target it compares clean/protected manipulation outputs using semantic, depth, scene-structure and intermediate-feature discrepancies. With semantic/depth weights, scene embedding, target-layer features and mask resizing defined as below, the four terms are:

$$\mathcal{L}^{(k)}_{\text{sem}} = -\left\| \left(M_k(x) - M_k(x_p) \right) \odot m_{\text{sem}} \right\|_2^2, \tag{2}$$

$$\mathcal{L}^{(k)}_{\text{depth}} = -\left\| \left(M_k(x) - M_k(x_p) \right) \odot m_{\text{depth}} \right\|_2^2, \tag{3}$$

$$\mathcal{L}^{(k)}_{\text{struct}} = -\left[\left(1 - \text{SSIM}\left(M_k(x), M_k(x_p) \right) \right) + \left(1 - \cos\left(S(M_k(x)), S\left(M_k(x_p) \right) \right) \right) \right], \tag{4}$$

$$\mathcal{L}^{(k)}_{\text{feat}} = -\sum_{l \in \{1,2,3\}} \left\| \left(E^{(k)}_l(x) - E^{(k)}_l(x_p) \right) \odot B_l(m_{\text{sem}}) \right\|_2^2. \tag{5}$$

The per-target SSDM loss sums these terms. Depth values are clipped away from zero before inverse-depth normalization. Target-specific losses are combined by the stabilized KPI weighting in Eqs. (6)–(8), adapted from ID-Guard [2].

$$\kappa^t_k = \text{clip}(\text{KPI}^t_k, \varepsilon_w, 1 - \varepsilon_w), \tag{6}$$

$$\beta^{t+1}_k = \varepsilon_w - (1 - \kappa^t_k)\ln(\kappa^t_k), \tag{7}$$

$$w^t_k = \frac{\alpha_k \beta^t_k}{\sum_j \alpha_j \beta^t_j}. \tag{8}$$

Clipping removes endpoint singularities and keeps normalized multi-target weights well defined.

4.3 Auxiliary JPEG-Aware Forensic Module

JAFM is an auxiliary protocol-characterization branch, not a source of the main proactive claim. It combines an ImageNet-pretrained ResNet18 image branch with a JPEG DCT-domain branch for single-versus-double JPEG classification. When used in CRPS, its fused encoder representation is frozen and serves only as a forensic feature embedding.

The reported JAFM checkpoint is trained for 50 epochs. Images are decoded, resized to 224×224 by bilinear interpolation, geometrically augmented only in training, and losslessly re-encoded to PNG before JPEG synthesis. Single samples follow PNG→JPEG(q2); double samples follow PNG→JPEG(q1)→JPEG(q2), so both classes share the same final quality q2.

Using seed 42 and val_ratio=0.1 gives 11,353 training and 1,261 validation source images; no independent source-disjoint test split is used. Training samples 15 QF pairs. Full-corpus protocol evaluation applies 50/75, 50/85, 75/85, 75/95 and 85/95 to all 12,614 sources, yielding 126,140 samples. Table 3 therefore characterizes this protocol rather than held-out-source generalization.

4.4 Compression-Robust Perturbation Strategy

CRPS uses a smooth periodic surrogate for JPEG's non-differentiable rounding, following differentiable-JPEG practice [9]. For normalized coefficient z, Eq. (9) defines the surrogate:

$$\mathrm{softRound}(u) = u - \frac{\sin(2\pi u)}{2\pi}, \qquad \tilde{f} = q\,\mathrm{softRound}(f/q). \tag{9}$$

For sampled JPEG quality q, Eq. (10) penalizes changes in the weighted SSDM objective after recompression:

$$\mathcal{L}_{\mathrm{CRPS}} = \mathbb{E}_Q \left| \sum_k w_k \mathcal{L}_{\mathrm{SSDM}}^{(k)}(x, x_p) - \sum_k w_k \mathcal{L}_{\mathrm{SSDM}}^{(k)}\left(x, J_Q(x_p)\right) \right|. \tag{10}$$

The optional forensic regularizer in Eq. (11) uses the frozen JAFM encoder to separate conventional and protected double-compression histories:

$$\mathcal{L}_{\mathrm{for}} = -\left\| F_{\mathrm{JAFM}}\left(J_{Q_2} \circ J_{Q_1}(x)\right) - F_{\mathrm{JAFM}}\left(J_{Q_2} \circ J_{Q_1}(x_p)\right) \right\|_1, \qquad Q_1 \neq Q_2. \tag{11}$$

Matched ablations enable JPEG exposure and the auxiliary terms separately to isolate recompression effects. CRPS is therefore a mechanism/robustness extension, not part of the primary SSDM checkpoint (Fig. 5).

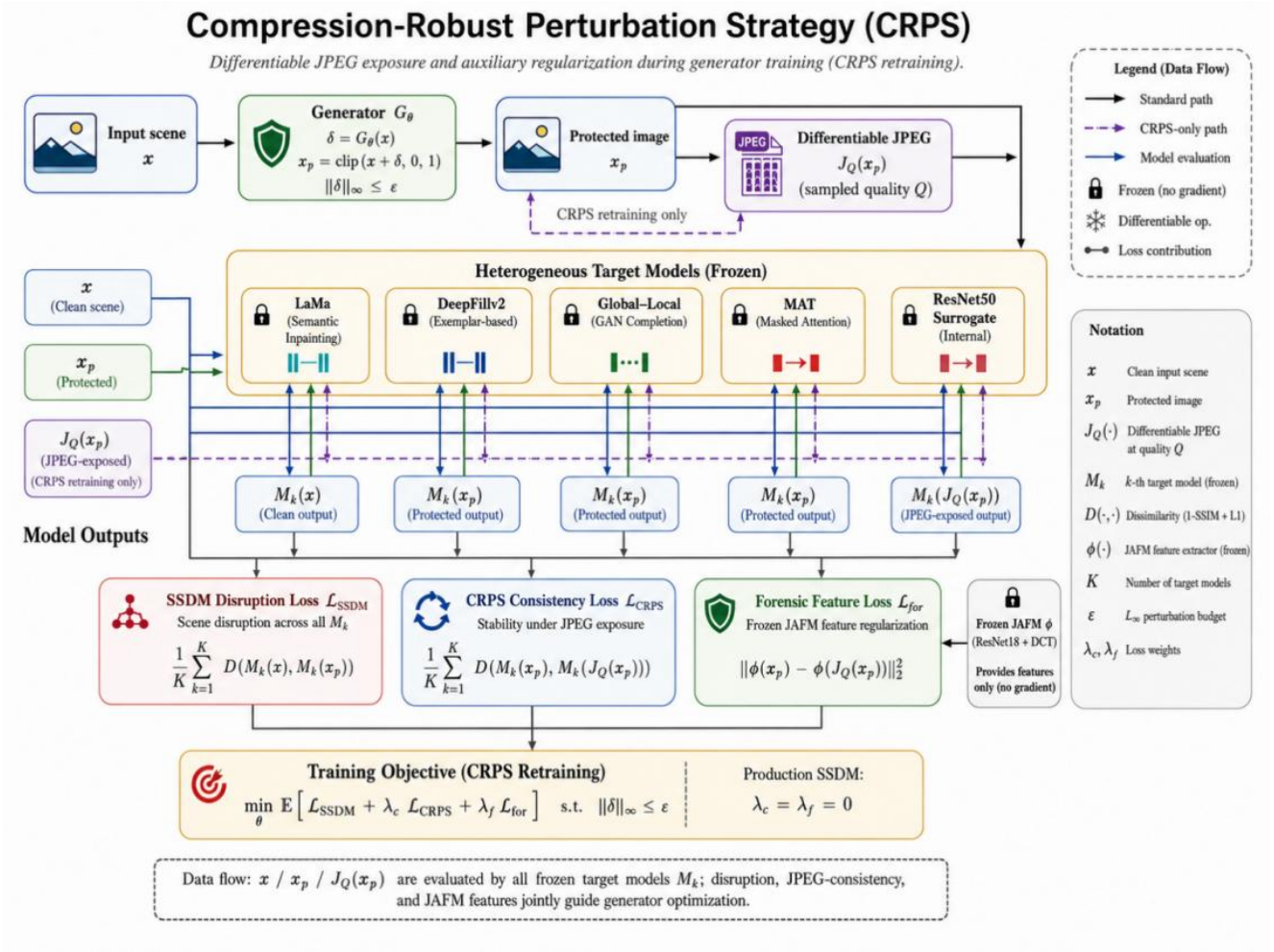


Fig. 5 CRPS matched compression-aware extension. Differentiable JPEG exposes protected images to sampled qualities; consistency compares weighted SSDM objectives before/after recompression, and optional frozen-JAFM features regularize double-compression representations. The primary SSDM checkpoint uses λc=λf=0.

4.5 Datasets and target models

ADE20K supplies 22,210 training images across 150 scene categories [10]. CASIA v2 supplies 12,614 proactive-evaluation source images (7,491 authentic; 5,123 tampered) [11]. For SSDM, CASIA is a source-scene corpus rather than an authenticity benchmark: labels are unused in optimization and SDR, and all images undergo the same protected-versus-clean inpainting/completion comparison. The target panel comprises LaMa [12], DeepFillv2 [13], Global–Local [14], MAT [15], and an internal ResNet50 encoder-decoder surrogate. The surrogate has an ImageNet-V2-initialized ResNet50 encoder and randomly initialized decoder, receives no task-specific training, and is frozen; it is therefore only an auxiliary architectural target, not a task-validated editor.

For proactive SSDM experiments, ADE20K and CASIA images are converted to RGB, resized to 256×256 by bilinear interpolation without aspect-ratio preservation or cropping, and scaled to [0,1]. A fixed centered 128×128 square mask (25% image area) is applied identically to every target and to clean/protected paths. All targets receive the same tensor and composite outputs as x⊙(1−m)+out⊙m at 256×256, so paired outputs are aligned by construction. SDR is 1−SSIM using an 11×11 Gaussian window (σ=1.5) with no further alignment; PSNR compares source and protected images at the same resolution. JAFM preprocessing is described in Section 4.3.

4.6 Training and implementation

The primary SSDM generator is trained at 256×256 resolution for 15 epochs with batch size 16 using Adam. The learning rate is fixed at 10^{-4} after a three-epoch linear warm-up. The perturbation budget is ε=0.08 under the L∞ norm, with λ_adv=25, λ_struct=8, λ_mask=0.1 and λ_mag=5. A short PGD warm start against LaMa can initialize the scene prior; test-time refinement uses ten projected-gradient steps at the same budget. This setting is a proof-of-concept point on the disruption–fidelity trade-off. Compression-aware variants use matched 15-epoch retraining and are evaluated at JPEG qualities 50, 75 and 90.

4.7 Evaluation metrics and statistical reporting

SDR (Eq. 12) is 1−SSIM between manipulation outputs from protected and clean inputs. It directly measures protected-versus-clean structural output divergence, not whether an intended edit succeeds or fails:

$$\mathrm{SDR} = 1 - \mathrm{SSIM}\left(M(x_p), M(x)\right). \tag{12}$$

SemDR (Eq. 13) is one minus cosine similarity between scene embeddings of the two manipulation outputs. For target M_k and scene encoder S:

$$\mathrm{SemDR}_k = 1 - \cos\left(S(M_k(x)), S\left(M_k(x_p)\right)\right). \tag{13}$$

SemDR is reported for component ablation and the paired generic-perturbation control, not Table 1. In matched CRPS ablation, compression survival (CS) is the mean per-sample ratio of post-JPEG SDR to the corresponding clean protected-image SDR, distinct from a ratio of aggregate means. Fidelity uses PSNR/SSIM; JAFM uses accuracy, precision, recall and F1. SDR comparisons are therefore output-divergence comparisons, not edit-prevention, randomness or uncertainty measures. The ε=0.08 point is interpreted jointly with fidelity. A separate 2,000-image control run at the same ε=0.08 compares Scene-Guard with uniform, Gaussian-matched and sign perturbations under the fixed-mask five-target protocol. All methods in that run share the same target instances, including the same randomly initialized auxiliary decoder, so the comparison is paired and interpreted only within that run rather than pooled with Table 1. Supplementary Table S7 reports mean SDR, across-target SD, SemDR, PSNR and SSIM.

Reference methods are No Defense, CMUA-Watermark [16], PhotoGuard [3], IAP [17] and adapted ID-Guard [2]. This unified adaptation protocol is a reference, not a reproduction of published operating points or a state-of-the-art ranking. No pretrained defense checkpoints are loaded. Each perturbation uses 20-step PGD at ε=0.08 and step size 0.01; CMUA and ID-Guard optimize all five frozen targets (equal-weight and dynamic-weight MSE), whereas PhotoGuard and IAP optimize LaMa only (VGG-encoder MSE; MSE+smoothness). None uses a mask. This optimizer is distinct from Scene-Guard's separate 10-step refinement. The 2,000-image subset is non-random with no sampling seed; Table 1 SD is across target means, not run-to-run uncertainty. Full settings are in Supplementary Table S2a.

4.8 AI-assisted manuscript preparation

OpenAI ChatGPT assisted English editing, organization, table formatting, figure-layout suggestions and consistency checking, but not research images, experimental data, analyses or scientific results. All suggestions were reviewed by the authors, who retain responsibility for the manuscript.
Supplementary information. Supplementary Figs. S1–S3 and Tables S1, S2, S2a and S3–S7 are provided separately.

Declarations

Funding. No specific funding was received.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable; only public image datasets were used.

Consent for publication. Not applicable.

Data availability. ADE20K and CASIA v2 are publicly available and cited. Processed splits, evaluation lists and metadata are available from the corresponding author on reasonable request.

Materials availability. Not applicable.

Code availability. Source code, checkpoints and evaluation scripts are available from the corresponding author on reasonable request.

Author contributions. Conceptualization and methodology: Min Shi, Yingnan Zhang; software: Min Shi, Hemin Yin, Yanzhe Zhang; validation: Hemin Yin, Yanzhe Zhang; formal analysis: Min Shi, Yingnan Zhang, Hemin Yin; investigation: Min Shi, Hemin Yin, Yanzhe Zhang; data curation: Hemin Yin, Yanzhe Zhang; visualization: Yanzhe Zhang; writing—original draft: Min Shi; writing—review and editing: Min Shi, Yingnan Zhang, Hemin Yin, Yanzhe Zhang. All authors approved the manuscript.

References

[1] Yang, P., Ni, R. & Zhao, Y. Double JPEG compression detection by exploring the correlations in DCT domain. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (IEEE, 2018).

[2] Qu, Z., Lu, W., Luo, X., Wang, Q. & Cao, X. ID-Guard: A universal framework for combating facial manipulation via breaking identification. IEEE Trans. Pattern Anal. Mach. Intell. 48, 1720–1735 (2026). https://doi.org/10.1109/TPAMI.2025.3616232.

[3] Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A. & Madry, A. Raising the cost of malicious AI-powered image editing. In Proceedings of the 40th International Conference on Machine Learning 29894–29918 (2023).

[4] Chen, R. et al. EditShield: Protecting unauthorized image editing by instruction-guided diffusion models. In Computer Vision – ECCV 2024, Lecture Notes in Computer Science 15121, 126–142 (Springer, 2024). https://doi.org/10.1007/978-3-031-73036-8_8.

[5] Sandoval-Segura, P., Geiping, J. & Goldstein, T. JPEG compressed images can bypass protections against AI editing. arXiv:2304.02234 (2023).

[6] Kirillov, A., Girshick, R., He, K. & Dollar, P. Panoptic feature pyramid networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 6399–6408 (2019).

[7] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. & Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. 44, 1623–1637 (2022).

[8] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464 (2018).

[9] Reich, C., Debnath, B., Patel, D. & Chakradhar, S. Differentiable JPEG: The devil is in the details. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 4126–4137 (2024).

[10] Zhou, B. et al. Scene parsing through ADE20K dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 633–641 (2017).

[11] Dong, J., Wang, W. & Tan, T. CASIA image tampering detection evaluation database. In IEEE China Summit and International Conference on Signal and Information Processing 422–426 (2013).

[12] Suvorov, R. et al. Resolution-robust large mask inpainting with Fourier convolutions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2149–2159 (2022).

[13] Yu, J. et al. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision 4471–4480 (2019).

[14] Iizuka, S., Simo-Serra, E. & Ishikawa, H. Globally and locally consistent image completion. ACM Trans. Graph. 36, Article 107, 107:1–107:14 (2017). https://doi.org/10.1145/3072959.3073659.

[15] Li, W. et al. MAT: Mask-aware transformer for large hole image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 10758–10768 (2022).

[16] Huang, H. et al. CMUA-Watermark: A cross-model universal adversarial watermark for combating deepfakes. Proc. AAAI Conf. Artif. Intell. 36, 989–997 (2022). https://doi.org/10.1609/aaai.v36i1.19982.

[17] Zhu, Y. et al. Information-containing adversarial perturbation for combating facial manipulation systems. IEEE Trans. Inf. Forensics Secur. 18, 2046–2059 (2023). https://doi.org/10.1109/TIFS.2023.3262156.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx