Robust Reversible Recovery for Adversarially Protected JPEG Images

Dahao Fu
School of Cyber Science and Technology
Sun Yat-sen University
Shenzhen, China
fudahao2021@163.com

Jiangqun Ni*
School of Cyber Science and Technology
Sun Yat-sen University
Shenzhen, China
issjqni@mail.sysu.edu.cn

Abstract

Reversible adversarial examples (RAEs) can disrupt access by malicious AI models while ensuring recoverability for authorized users. However, practical image circulation is dominated by JPEG files, whereas most existing RAE methods perturb spatial-domain images and are therefore poorly matched to compression, re-saving, and coefficient quantization. Beyond JPEG compatibility, circulation imposes a stricter and underexplored requirement that protected images should be perturbed in the JPEG domain, remain adversarial after spatial reconstruction, preserve visual fidelity, incur modest file-size growth, and remain recoverable under channel attacks such as recompression, noise, cropping, resizing, and platform re-saving. This paper presents the first systematic study of robust reversible adversarial protection for post-processed JPEG images and proposes **SRAP-JPEG**, a Synchronized Reversible Adversarial Protection framework on quantized DCT coefficients. SRAP-JPEG decouples adversarial perturbation from reversible recording by injecting perturbations into luminance coefficients, hiding recovery records in paired chrominance coefficients through a secret-key mapping, and adopting coefficient-adaptive allocation to trade off reversible capacity, attack strength, visual quality, and file-size growth. It crafts protective perturbations directly in the JPEG coefficient domain by propagating gradients from classification or vision-language objectives through the chain rule. For authorized recovery, SRAP-JPEG integrates block-phase synchronization, embedding-state identification, and coefficient restoration, enabling exact recovery of intact protected JPEGs and high-fidelity recovery after practical distortions. Experiments on ImageNet and MS-COCO show that SRAP-JPEG achieves an 81.82% average attack success rate on ten mainstream classifiers, degrades CLIP-based image-text retrieval, and recovers distorted protected images with high visual fidelity, including 41.60 dB PSNR after JPEG recompression at QF 70 and 59.51 dB on the aligned uncropped region after edge cropping at ratio 0.25.

1 Introduction

Large-scale visual data has become a key driver of deep model development [1, 2]. Once shared online, however, images could be collected by unauthorized parties for model training, content analysis, identity inference, or cross-modal retrieval. Adversarial examples (AEs) could be used to protect the images by effectively preventing unauthorized analysis by malicious AI models [3, 4], although the conventional AEs are irreversible, and the protected images remain degraded even for authorized users, which is not allowed for lossless applications, e.g., military reconnaissance, medical

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

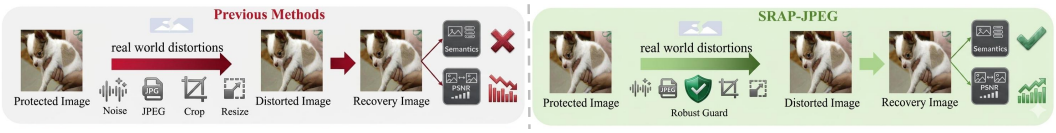

Figure 1: Motivation of SRAP-JPEG. The practical lossy channels could disrupt the adversarial protection and corrupt reversible records; SRAP-JPEG targets robust reversibility for JPEG images for lossy channels.

imaging and remote sensing. Reversible adversarial examples (RAEs) tackle this by incorporating adversarial perturbations with reversible data hiding (RDH), such that the protected images remain adversarial before authentication and could be recovered for authorized users [5, 6].

Existing RAE methods achieve these dual objectives through RDH-based embedding, generative restoration, and JPEG-domain coefficient manipulation [7–10]. For practical applications, the JPEG RAE is of great importance, as the shared images are usually stored, transmitted in JPEG compressed form. However, JPEG compatibility alone does not ensure robust recovery. Most existing RAEs define reversibility for lossless channels, where perturbations and location maps could be correctly extracted before restoration. This assumption is no longer the case for practical lossy channels, where the channel attacks, e.g., recompression, noise, cropping and resizing as shown in Fig. 1, could alter these records and break the perturbation-record correspondence required for recovery. Therefore, the reversibility over lossless channels as conventionally understood is by no means the robust reversibility, in other words, an RAE image that is exactly recoverable for lossless channels is likely to become unrecoverable for lossy channels. This gap motivates the development of robust reversible adversarial protection against channel attacks as a distinct design objective.

The greater challenge lies in the large payload required for reversible adversarial protection. To disrupt malicious AI models, perturbations must be applied to a substantial number of pixels or JPEG coefficients. Recovering the protected image therefore requires perturbation records that can survive lossy-channel distortions. This makes robust reversible adversarial protection more demanding than conventional robust watermarking, which usually embeds a smaller payload [11], and also different from image restoration, which recovers image content from degraded observations rather than preserving exact perturbation records [12]. For JPEG images, reversibility is even more difficult because coefficient modification must account for quantization, file-size expansion, compression compatibility, and limited reversible capacity [13–15]. We therefore formulate robust reversibility as follows: protected images should remain adversarially effective before authentication, while authorized recovery should remove adversarial effects, exactly recover the original image under lossless channels, and restore the image with the highest possible fidelity under lossy channels.

We propose **SRAP-JPEG**, a unified JPEG-domain framework for robust reversible adversarial protection over lossy channels. SRAP-JPEG alleviates the limited allocatable capacity of JPEG images by separating adversarial perturbations from recovery records: perturbations are crafted in luminance coefficients, whereas recovery records are embedded into paired chrominance coefficients through a key-controlled mapping that links perturbations to records and enables key-triggered recovery. A coefficient-adaptive allocation scheme further selects valid pairs and embedding strength according to image statistics, balancing reversible payload, perceptual quality, file-size growth, and attack strength. By propagating gradients through the JPEG-to-spatial transformation, SRAP-JPEG generates JPEG-domain perturbations that mislead spatial-domain AI models, while distortion-aware recovery restores lossy-channel images through block-phase synchronization, embedding-state identification, and coefficient restoration.

Our main contributions are four-fold: (1) We formulate and systematically study robust reversible adversarial protection for JPEG images over lossy channels. (2) We develop a key-controlled cross-channel perturbation-recovery mechanism that decouples adversarial perturbations in luminance coefficients from reversible records in chrominance coefficients, thereby establishing perturbation-record correspondence for authorized recovery. (3) We devise a coefficient-adaptive allocation scheme with JPEG-domain adversarial optimization that balances perturbation strength, visual quality, file-size growth, and reversible capacity. (4) Experiments on ImageNet and MS-COCO demonstrate adversarial effectiveness against classifiers and CLIP-based retrieval, while maintaining high perceptual quality, controlled file-size growth, and reliable recovery under channel attacks such

as recompression, noise, cropping, and resizing. Specifically, SRAP-JPEG achieves an average attack success rate of 81.82% on ten ImageNet classifiers and recovers distorted protected images with high fidelity, e.g., 41.60 dB PSNR under JPEG recompression at QF 70 and 59.51 dB under edge cropping at ratio 0.25.

2 Related work

Reversible adversarial examples (RAEs) protect images by making them adversarial before authentication while allowing authorized users to recover the original images. Existing RAEs mainly achieve reversibility through RDH-style embedding of perturbations or auxiliary records [6, 16, 7, 8, 17], reversible transformations, spatial reorganization, visible perturbation structures [18–24], and learning-based adversarial generation and restoration [10]. Recent studies further extend RAEs to feature-space dataset protection, traceable protection, label-free encoder attacks, JPEG-domain protection [25, 26, 9, 27], improving their applicability to different protection scenarios. These methods show that both the adversarial protection and authorized recovery could be achieved for lossless channels. As a result, the task of reversible recovery becomes challenging for lossy channels, where recompression, noise, cropping, and resizing can disturb both the adversarial perturbations and the auxiliary recovery information.

One class of relevant studies are robust and reversible watermarking as they also target at message extraction and cover recovery for lossy and lossless channels. Classical robust watermarking, including spread-spectrum watermarking, aims to detect or extract compact messages under signal processing or geometric changes [11]; reversible watermarking further requires cover recovery and exposes the trade-off among robustness, reversibility, capacity, and perceptual quality [13, 14, 5].In addition, their JPEG domain variants should also manage DCT coefficient selection, quantization behavior, and file-size expansion [15]. Although these constraints are relevant to SRAP-JPEG, the adversarial protection is more demanding because it must preserve high-dimensional semantic perturbation records rather than a small ownership or authentication payload.

Image restoration can recover high-fidelity visual content from degraded observations [12], but its objective is fundamentally different from robust authorized recovery, as visually restored images may still remain adversarially effective. In adversarial protection, recovery must remove adversarial semantic behavior and reconstruct the authorized JPEG content from protected records after possible channel attacks, rather than merely recover a visually faithful image. SRAP-JPEG is therefore positioned as a high-payload robust reversible adversarial protection framework for JPEG images, rather than a conventional RAE, watermarking, or image restoration method.

3 Method

3.1 Overview

We propose **SRAP-JPEG**, a **S**ynchronized **R**eversible **A**dversarial **P**rotection framework for JPEG images. Given a cover JPEG image, SRAP-JPEG generates a protected JPEG that remains adversarial to downstream models, preserves visual quality and controlled file-size variation, and supports reversible recovery for lossy channels or format conversion. This setting reflects normal image circulation, where the protected file may be re-saved, recompressed, or converted before authorization.

As illustrated in Fig. 2 and Fig. 3, SRAP-JPEG has two stages: **adversarial protection** and **distortion-aware reversible recovery**. In the protection stage, the cover JPEG is represented by quantized DCT coefficients, where Y-channel coefficients carry adversarial perturbations and paired Cb/Cr coefficients record the recovery information. A secret-key-controlled mapping links each perturbation position with its recovery position. Coefficient-adaptive allocation chooses suitable coefficient pairs and embedding strength, while JPEG-domain optimization updates the perturbations according to downstream model objectives. In the recovery stage, SRAP-JPEG combines block-phase synchronization, embedding-state identification, and coefficient restoration to recover authorized content without assuming access to an intact compressed bitstream. This design keeps adversarial perturbation generation and reversible recording synchronized within the JPEG representation.

The following subsections describe these components in order. Sec. 3.2 introduces keyed cross-channel perturbation–recovery coupling; Sec. 3.3 presents coefficient-adaptive capacity allocation;

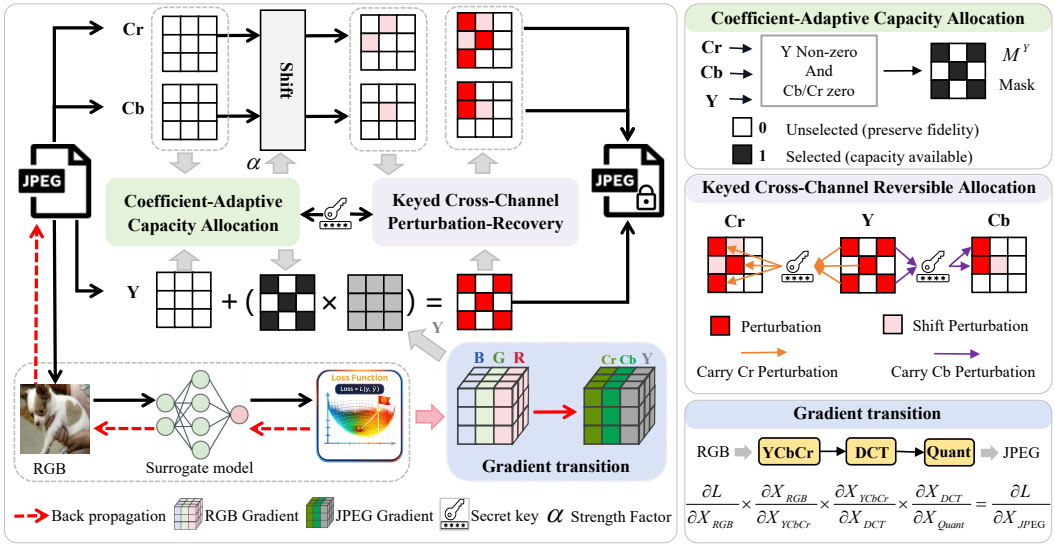

Figure 2: Overview of SRAP-JPEG. Adversarial perturbations are applied to Y-channel coefficients, while reversible records are embedded into paired Cb/Cr coefficients through a secret-key mapping. Gradients are propagated from the RGB domain to the JPEG domain for coefficient-level optimization.

Sec. 3.4 details JPEG-domain adversarial optimization; and Sec. 3.5 describes distortion-aware reversible recovery.

3.2 Keyed Cross-Channel Perturbation-Recovery Coupling

Given an input image $\mathbf{X}$, we write the JPEG processing pipeline as $\mathbf{X} \to \mathbf{X}_{\mathrm{YCbCr}} \to \mathbf{X}_{\mathrm{DCT}} \to \mathbf{X}_{\mathrm{Quant}} \equiv \mathbf{X}_{\mathrm{JPEG}}$, where the terms denote color-space, block-DCT, and quantized coefficient representations. We write $\mathbf{X}_{\mathrm{JPEG}} = \{\mathbf{X}^{Y}_{\mathrm{JPEG}}, \mathbf{X}^{Cb}_{\mathrm{JPEG}}, \mathbf{X}^{Cr}_{\mathrm{JPEG}}\}$ for its three channels.

SRAP-JPEG decouples adversarial perturbation and reversible recording across these channels. Y-channel coefficients carry adversarial perturbations, while Cb/Cr coefficients store the reversible records required for authorized recovery. This separation lets perturbation optimization and recovery recording be controlled independently.

To securely bind perturbation and recovery, we introduce a secret-key-controlled cross-channel mapping. Let $\mathcal{P}_Y$ contain 26 Y-channel DCT positions and $\mathcal{P}_C$ contain 13 positions from each of Cb and Cr, yielding 26 channel-tagged slots. These fixed, public sets are selected using the standard JPEG quantization tables (Appendix D); the secret key K specifies a bijection $\pi_K : \mathcal{P}_Y \to \mathcal{P}_C$. Each perturbation at p is associated with a unique chrominance position $q = \pi_K(p)$, where the reversible record is embedded. During decoding, the same key reproduces this correspondence, enabling records to be mapped back to their original Y-channel positions while making the perturbation–recovery relation difficult to infer without K.

3.3 Coefficient-Adaptive Capacity Allocation

Not all JPEG coefficients are suitable for perturbation and reversible recording. For each candidate pair (p, q) with $q = \pi_K(p)$, validity requires $\mathbf{X}^{Y}_{\mathrm{JPEG}}(p) \neq 0$ and $\mathbf{X}^{Cb/Cr}_{\mathrm{JPEG}}(q) = 0$, where $\mathbf{X}^{Cb/Cr}_{\mathrm{JPEG}}$ denotes either chrominance channel. The nonzero Y coefficient avoids file-size inflation, while the zero Cb/Cr coefficient provides a clean embedding state. These constraints make the valid locations image-dependent, motivating adaptive capacity allocation.

We define the binary mask for valid Y-channel perturbation locations as

$$\mathbf{M}^{Y}(p) = \begin{cases} 1, & \mathbf{X}^{Y}_{\mathrm{JPEG}}(p) \neq 0 \text{ and } \mathbf{X}^{Cb/Cr}_{\mathrm{JPEG}}(\pi_K(p)) = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

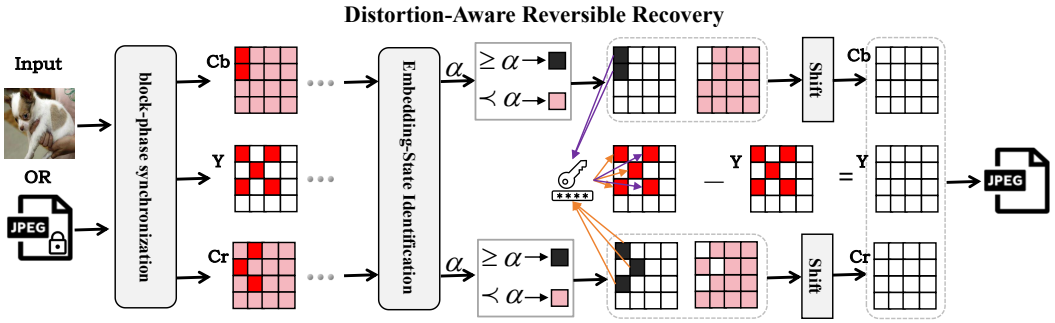

Figure 3: Distortion-aware recovery pipeline. Block-phase synchronization and embedding-state identification locate the recovery state, after which Cb/Cr records restore the JPEG coefficients.

The number of valid locations, adaptive embedding strength, and chrominance coefficient shifting are defined jointly as

$$N = \sum_{p} \mathbf{M}^{Y}(p), \qquad \alpha = \begin{cases} 2, & N < \tau, \\ 1, & N \geq \tau, \end{cases} \qquad \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr} = \begin{cases} \mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr} + \alpha, & \mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr} > 0, \\ \mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr} - \alpha, & \mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr} < 0, \\ 0, & \mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr} = 0. \end{cases} \tag{2}$$

where τ is a predefined threshold and $\tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}$ is the shifted Cb/Cr coefficient. When valid capacity is limited, $\alpha = 2$ enlarges the recordable perturbation range; otherwise, $\alpha = 1$ favors visual quality and storage efficiency. The shifting moves original nonzero chrominance coefficients outside $[-\alpha, \alpha]$, so values inside this interval can be interpreted as embedded records during later recovery.

3.4 Gradient-Driven JPEG-Domain Adversarial Optimization

After establishing the valid cross-channel pairs and embedding strength α, SRAP-JPEG optimizes adversarial perturbations directly in the JPEG domain while synchronously recording reversible perturbation information. The adversarial objective depends on the protected task and is written compactly as

$$\mathbb{L}_{\mathrm{adv}} = \begin{cases} \mathrm{CE}\left(\sum_{k=1}^{L} w_k l_k(\mathbf{x}),\, y\right), & \text{classification model,} \\ -\dfrac{f_I(\mathbf{x})^{\top} f_T(\mathbf{t})}{\|f_I(\mathbf{x})\|_2 \, \|f_T(\mathbf{t})\|_2}, & \text{vision-language model,} \end{cases} \tag{3}$$

where l_k, w_k, and L denote surrogate logits, fusion weights, and the number of classifiers, while f_I and f_T denote image and text encoders. SRAP-JPEG performs projected signed-gradient ascent on L_{adv}: maximizing cross-entropy reduces the true-label confidence for classifiers, while maximizing the negative cosine similarity reduces image-text alignment for vision-language models. The detailed objective construction is provided in the Appendix.

Although L_{adv} is evaluated in the RGB domain, perturbations are applied to quantized JPEG coefficients. Thus, gradients are propagated through the JPEG processing chain $\mathbf{X} \to \mathbf{X}_{\mathrm{YCbCr}} \to \mathbf{X}_{\mathrm{DCT}} \to \mathbf{X}_{\mathrm{Quant}} \to \mathbf{X}_{\mathrm{JPEG}}$:

$$\frac{\partial L_{\mathrm{adv}}}{\partial \mathbf{X}_{\mathrm{JPEG}}} = \frac{\partial L_{\mathrm{adv}}}{\partial \mathbf{X}} \cdot \frac{\partial \mathbf{X}}{\partial \mathbf{X}_{\mathrm{YCbCr}}} \cdot \frac{\partial \mathbf{X}_{\mathrm{YCbCr}}}{\partial \mathbf{X}_{\mathrm{DCT}}} \cdot \frac{\partial \mathbf{X}_{\mathrm{DCT}}}{\partial \mathbf{X}_{\mathrm{Quant}}}. \tag{4}$$

The color conversion and block-wise DCT are linear transforms. Although standard JPEG quantization contains non-differentiable rounding, SRAP-JPEG optimizes the quantized coefficient representation directly, so the gradient path only requires quantization-table scaling; Appendix B provides the detailed argument.

Let $\delta_Y^{(t)}$ denote the Y-channel coefficient perturbation at iteration t, initialized as $\delta_Y^{(0)} = \mathbf{0}$. To avoid file-size inflation and preserve reversible capacity, updates are restricted to valid locations by $\mathbf{M}^Y$

and projected to the reversible range $[-\alpha, \alpha]$:

$$\delta_Y^{(t+1)} = \mathrm{clip}\left(\delta_Y^{(t)} + \mathbf{M}^Y \odot \mathrm{sign}\left(\frac{\partial L_{\mathrm{adv}}}{\partial \mathbf{X}_{\mathrm{JPEG}}^{Y}{}^{(t)}}\right), -\alpha, \alpha\right), \qquad \delta_{Cb/Cr}^{(t+1)}(\pi_K(p)) = \delta_Y^{(t+1)}(p). \tag{5}$$

For each valid keyed pair (p, q) with $q = \pi_K(p)$, the updated luminance coefficient and the paired chrominance record are

$$\mathbf{X}_{\mathrm{JPEG}}^{Y}{}^{(t+1)} = \mathbf{X}_{\mathrm{JPEG}}^{Y}{}^{(0)} + \delta_Y^{(t+1)}, \qquad \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr(t+1)} = \tilde{\mathbf{x}}_{\mathrm{JPEG}}{}^{(0)} + \delta_{Cb/Cr}^{(t+1)}(\pi_K(p)). \tag{6}$$

The paired chrominance coefficients are updated with these records inside the reserved interval, and the resulting Y, Cb, and Cr coefficients are encoded as the protected JPEG image.

3.5 Distortion-Aware Reversible Recovery

In practical circulation, the protected image may be recompressed, resized, cropped, or converted to another format. These operations can shift or disrupt the original JPEG block grid, so recovery cannot simply invert coefficients at fixed positions. SRAP-JPEG therefore performs distortion-aware reversible recovery in three stages: block-phase synchronization, embedding-state identification, and perturbation extraction with coefficient restoration. This design keeps recovery independent of an intact compressed bitstream and only assumes the shared secret key used for cross-channel pairing.

Block-phase synchronization. Given a received image $\tilde{\mathbf{x}} \in \mathbb{R}^{3\times h\times w}$, we first estimate the correct 8×8 block alignment. We enumerate offsets (i, j) with $i, j \in \{0, \ldots, 7\}$, extract the corresponding aligned blocks, and compute their DCT coefficients $\mathbf{x}_{\mathrm{DCT}}^{(i,j)}$. The phase is selected by integer-quantization consistency:

$$\mathcal{D}(i,j) = \sum_b \left\| \frac{\mathbf{x}_{\mathrm{DCT},b}^{(i,j)}}{\mathbf{Q}} - \mathrm{round}\left(\frac{\mathbf{x}_{\mathrm{DCT},b}^{(i,j)}}{\mathbf{Q}}\right) \right\|_1, \qquad (i_s, j_s) \quad = \arg\min_{i,j} \mathcal{D}(i,j), \tag{7}$$

where b indexes 8×8 blocks and $\mathbf{Q}$ denotes the JPEG quantization table used for the corresponding channel. The estimated offset (i_s, j_s) is used to extract the largest spatially aligned sub-image and compute its JPEG-domain representation for recovery. Quantization consistency can be ambiguous after non-aligned recompression because the new JPEG grid competes with the original embedding grid (Appendix M).

Embedding-state identification. After synchronization, the decoder obtains an aligned JPEG representation $\tilde{\mathbf{x}}_{\mathrm{JPEG}}$, but the protection strength is unknown. We use an embedding-state analyzer to estimate $\hat{\alpha} \in \{0, 1, 2\}$, where $\hat{\alpha} = 0$ denotes no embedded perturbation and $\hat{\alpha} \in \{1, 2\}$ denotes the embedding strength. The decision relies on low-frequency chrominance DCT statistics: Sec. 3.3 shifts original nonzero Cb/Cr coefficients outside $[-\alpha, \alpha]$, while reversible records remain inside this interval. Thus, concentration near zero indicates records, and comparing proportions within $[-a, a]$ for $a \in \{1, 2\}$ determines the most plausible state. This compact test avoids extra side information about the selected strength while remaining compatible with distorted but statistically recognizable chrominance coefficients. Appendix C provides the histogram-template implementation.

Perturbation extraction and coefficient restoration. Given $\hat{\alpha}$, each keyed pair (p, q) with $q = \pi_K(p)$ is checked for a recoverable chrominance record:

$$\hat{\delta}_Y(p) = \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q), \quad \text{if } \left|\tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q)\right| \le \hat{\alpha}. \tag{8}$$

These records undo the Y-channel perturbation, and the chrominance coefficients are restored by reversing the shift:

$$\mathbf{x}_{\mathrm{JPEG}}^{Y}(p) = \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Y}(p) - \hat{\delta}_Y(p), \qquad \mathbf{x}_{\mathrm{JPEG}}^{Cb/Cr}(q) = \begin{cases} \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q) - \hat{\alpha}, & \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q) > \hat{\alpha}, \\ \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q) + \hat{\alpha}, & \tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q) < -\hat{\alpha}, \\ 0, & |\tilde{\mathbf{x}}_{\mathrm{JPEG}}^{Cb/Cr}(q)| \le \hat{\alpha}, \end{cases} \tag{9}$$

After restoring both luminance and chrominance channels, the recovered JPEG representation is converted back to the spatial domain to reconstruct the original image.

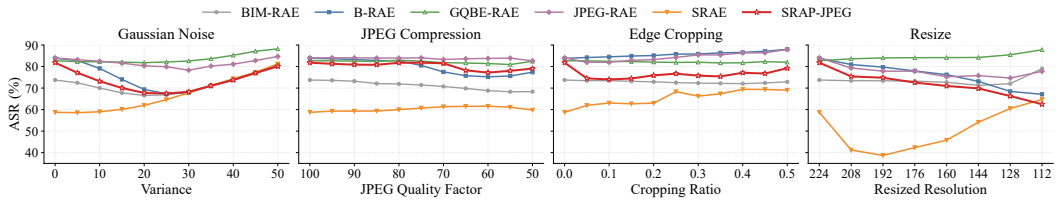

Figure 4: Attack robustness under common post-processing distortions. SRAP-JPEG maintains competitive ASR across Gaussian noise, JPEG recompression, edge cropping, and resizing.

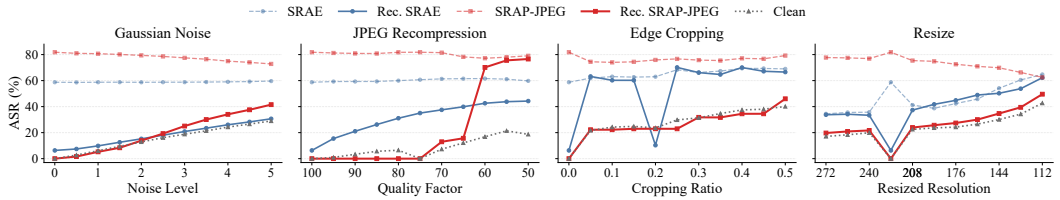

Figure 5: Semantic recovery robustness under Gaussian noise, JPEG recompression, edge cropping, and resizing. Each panel reports ASR before and after recovery; lower recovered ASR indicates more effective removal of adversarial behavior.

4 Experiments

4.1 Experimental Setup

We evaluate SRAP-JPEG on ILSVRC 2012 [28] and MS-COCO [29] using 1,000 images from each dataset. For reversible adversarial protection, we compare with five baselines: BIM-RAE [6], B-RAE [7], GQBE-RAE [8], SRAE [10], and JPEG-RAE [27]. For COCO multimodal attack evaluation, we additionally compare with AnyAttack [30]. All inputs are resized to 224×224 and encoded as JPEG images with quality factor 75 unless otherwise stated; the controlled experiments use 4:4:4 JPEG. Detailed baseline configurations, model lists, and implementation settings are provided in the Appendix. For ImageNet classification, we report attack success rate (ASR). For visual and recovery quality, we report PSNR and SSIM, and we use file expansion rate to measure JPEG compatibility. For COCO retrieval, we report text-retrieval and image-retrieval recall metrics, denoted by TR@K and IR@K. Detailed metric definitions are also provided in the Appendix D.

4.2 Robustness

Attack Robustness. Fig. 4 evaluates whether adversariality remains effective after common post-processing. SRAP-JPEG achieves 81.8% mean ASR on ImageNet, remaining close to the strongest reversible baselines despite operating under stricter constraints: perturbations must be embedded in JPEG coefficients, paired with recovery records, and kept compatible with later distortion-aware recovery. SRAP-JPEG remains robust to Gaussian noise, JPEG recompression, and cropping, with ASR generally staying in the high-70s to low-80s, and remains competitive under harder resizing, ranging from the mid-60s to 81.81%. These results show that SRAP-JPEG offers a practical trade-off between attack robustness and robust reversibility.

Recovery Robustness. Recovery robustness evaluates whether authorized recovery remains possible after common image-circulation operations, which is stricter than clean reversibility because the decoder must remove adversarial effects from recompressed, noisy, cropped, or resized protected JPEGs. Among existing RAE baselines, SRAE is the only method with limited distorted-image recovery ability; BIM-RAE, B-RAE, GQBE-RAE, and JPEG-RAE are designed for clean recovery and become fragile once the protected image is modified. We therefore compare recovery robustness with SRAE, while using the other baselines for attack ability, visual quality, and file-size comparisons. We report recovered ASR for semantic recovery and recovered-to-cover PSNR for visual recovery, where lower ASR and higher PSNR are better; for edge cropping, PSNR is computed only over the aligned uncropped region. Fig. 5 shows that SRAP-JPEG consistently restores semantic normality after distortion. Without distortion, SRAP-JPEG reduces ASR from 81.81% to 0%, while SRAE

Table 1: Compact recovery robustness summary distilled from the full sweeps in Appendix. All entries are PSNR (dB; higher is better), except that edge-cropping PSNR is computed only over the spatially aligned uncropped region. Each slash-separated tuple follows the listed settings; dashes indicate non-comparable crop compensation.

(a) Visual recovery: recovered-to-cover PSNR

Dist.	Settings	SRAE Rec.	**SRAP Rec.**	Clean-Dist.	Gain
Gaussian	0.5/2.5/5.0	38.34/31.19/25.95	80.01/35.24/28.24	45.39/32.10/26.18	**+41.67/+4.05/+2.29**
JPEG	95/70/50	32.46/27.06/25.71	∞/41.60/28.53	57.37/41.77/30.47	∞/**+14.54/+2.82**
Crop	0.05/0.25/0.50	23.18/22.75/22.40	59.43/59.51/59.79	$\infty/\infty/\infty$	**+36.25/+36.76/+37.39**
Resize	272/176/112	28.90/28.49/27.26	32.11/29.39/27.03	34.42/31.14/28.10	**+3.08/+1.23**/-0.23

(b) Distortion compensation: SRAP-JPEG under distorted protected inputs

Dist.	Settings	D-Prot./Prot.	**SRAP Rec.**	Clean-Dist.	Rec.-D-Prot.
Gaussian	0.5/2.5/5.0	42.77/31.87/26.14	80.01/35.24/28.24	45.39/32.10/26.18	**+37.24/+3.37/+2.10**
JPEG	95/70/50	52.23/39.37/28.24	∞/41.60/28.53	57.37/41.77/30.47	∞/**+2.23/+0.29**
Crop	0.05/0.25/0.50	$\infty/\infty/\infty$	59.43/59.51/59.79	$\infty/\infty/\infty$	–/–/–
Resize	272/176/112	32.70/28.79/27.21	32.11/29.39/27.03	34.42/31.14/28.10	-0.59/**+0.60**/-0.18

reduces ASR from 58.74% to 6.32%. This advantage remains under Gaussian noise, JPEG recompression, edge cropping, and resizing, indicating that the recovery records remain usable beyond clean-bitstream decoding. Cropping is especially challenging because it changes spatial support and disrupts JPEG block alignment; nevertheless, SRAP-JPEG achieves an average ASR reduction of 46.80 percentage points, far above the 6.16 points of SRAE.

Table 1 further evaluates visual recovery and distortion compensation. Panel (a) shows that SRAP-JPEG achieves substantially higher recovered-to-cover PSNR than SRAE across four common distortions and almost all tested settings. The gain is most pronounced under cropping, where SRAE provides little meaningful recovery but SRAP-JPEG restores images to nearly 60 dB. For protected JPEGs generated at $QF = 75$, exact recovery is observed in the tested aligned recompression settings with $QF \in \{75, 80, 85, 90, 95\}$ (Appendix Table 9). Additional double-JPEG tests with $QF_1 = 80$ show that strong recompression limits recovery (Appendix Table 15). Panel (b) reports positive Rec.-D-Prot. values under Gaussian noise and JPEG recompression, indicating that recovery not only removes adversarial perturbations but also compensates part of the external distortion. Overall, SRAP-JPEG is distinguished by robust authorized recovery rather than attack strength alone: it supports recovery after circulation in the evaluated settings, a challenging capability not provided by JPEG-RAE or most spatial-domain RAEs. Full sweeps and composed-circulation results are in Appendix J.

Real-platform circulation. We upsampled the same 50 protected images to 2048×2048 and uploaded each once to X, Facebook, and WeChat, yielding 150 upload-download instances. Across these workflows, protected ASR was 77.0%–77.5%, recovered ASR was 6.2%–6.8%, and recovered-to-cover PSNR was 39.86–41.20 dB. Appendix J reports the complete results and a separate 10-image study of upload resolution.

4.3 Attack Ability

After evaluating robust recovery, we examine whether SRAP-JPEG still preserves adversarial effectiveness under the constrained protection-recovery objective. Table 2 summarizes ImageNet classification and COCO retrieval results; complete metrics are reported in Appendix E Tables 5 and 7. For ImageNet, the displayed backbones are representative, and the Avg. row is computed over all 10 classifiers. SRAP-JPEG achieves 81.8% mean ASR, remaining comparable to the strongest reversible baselines despite stricter JPEG-domain constraints: perturbations are embedded in quantized coefficients, paired with keyed recovery records, kept under controlled file-size growth, and preserved for later distortion-aware recovery. This verifies that robust reversibility does not remove adversarial protection. For COCO retrieval, lower TR@Mean/IR@Mean indicates stronger disruption. Compared with AnyAttack [30], SRAP-JPEG lowers both retrieval means on all three OpenCLIP backbones and reduces their average from 47.2/46.8 to 28.1/23.1, showing that protected JPEGs also

Table 2: Transposed attack ability summary. Left: representative ImageNet ASR (↑). Right: COCO retrieval TR@Mean/IR@Mean (↓).

ImageNet ASR (%)↑							COCO TR@Mean/IR@Mean (%)↓			
Backbone	BIM-RAE	B-RAE	GQBE-RAE	SRAE	JPEG-RAE	**SRAP-JPEG**	Backbone	Clean	AnyAttack	**SRAP-JPEG**
RN-50	73.7	81.4	83.0	99.0	83.9	**84.6**	ViT-B/32	88.2/79.5	42.1/42.0	**19.0/16.5**
DN-121	80.1	72.6	88.6	32.8	90.1	**90.0**	ViT-B/16	88.4/79.6	47.0/47.1	**23.4/16.1**
Mob-v3	81.1	89.9	89.9	33.8	90.0	**90.3**	ViT-L/14	87.1/75.9	52.5/51.4	**41.8/36.8**
Avg.	73.7	83.8	82.7	58.7	84.0	**81.8**	Avg.	87.9/78.3	47.2/46.8	**28.1/23.1**

Table 3: Visual and recovery quality comparison. Each cell reports protected-image quality → recovered-image quality.

Metric	BIM-RAE	B-RAE	GQBE-RAE	SRAE	JPEG-RAE	**SRAP-ImageNet**	**SRAP-MSCOCO**
SSIM	$0.91 \rightarrow 1.00$	$0.85 \rightarrow 1.00$	$0.84 \rightarrow 0.99$	$0.68 \rightarrow 0.98$	$0.86 \rightarrow 1.00$	$0.85 \rightarrow 1.00$	$0.88 \rightarrow 1.00$
PSNR	$31.9 \rightarrow \infty$	$30.9 \rightarrow \infty$	$29.7 \rightarrow 46.1$	$24.4 \rightarrow 39.4$	$30.1 \rightarrow \infty$	$28.9 \rightarrow \infty$	$29.4 \rightarrow \infty$

disrupt image-text semantic matching. Feature-alignment and downstream poisoning analyses are provided in Appendix F and H.

4.4 Visual quality

Table 3 reports visual quality as protected-image quality followed by recovered-image quality. On the protection side, SRAP-JPEG keeps PSNR around 29 dB and SSIM between 0.85 and 0.88, which is slightly below BIM-RAE, B-RAE and JPEG-RAE because SRAP-JPEG must also reserve JPEG-domain capacity for keyed recovery records and distortion-aware recovery. This is a controlled trade-off under the stricter goal of jointly preserving adversarial protection, JPEG compatibility, robust reversibility, and security. On the recovery side, SRAP-JPEG exactly restores both ImageNet and MS-COCO protected images in the intact-JPEG setting, reaching SSIM 1.00 and PSNR ∞. Thus, the visual-quality cost mainly appears before authorization, while authorized users can recover the original image losslessly; qualitative comparisons are provided in Appendix Fig. 9.

4.5 Expansion Rate

File expansion is critical for deployment, as protected images must be stored, transmitted, and reused at scale. As shown in Fig. 6, SRAP-JPEG keeps storage overhead controlled: at the default quality factor Q=75, it produces 17.05 KB protected images, only 28% larger than 13.42 KB covers, while maintaining 28.92 dB PSNR, 0.85 SSIM, and 81.82% ASR. Across quality factors, the overhead remains 121.9%–131.9% despite the recovery records required for robust reversibility. SRAP-JPEG is also five to seven times smaller than spatial-domain RAE methods on average, supporting its practicality for large-scale JPEG datasets.

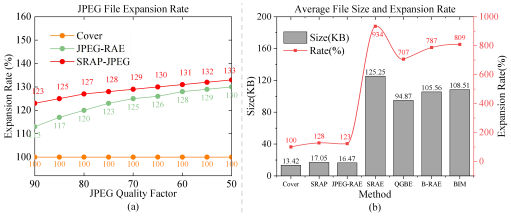

Figure 6: File expansion comparison.

4.6 Ablation Study

Table 4 examines whether each design component is necessary by reporting clean protection (ASR/PSNR) and recovery after Gaussian noise level 2.5, JPEG recompression with QF=70, crop ratio 0.25, and resize size 176; lower Rec. ASR and higher Rec. PSNR indicate better removal of adversarial effects and stronger visual restoration. The results show that the modules contribute complementary functions: adaptive allocation improves the attack-quality trade-off,

Table 4: Prot. reports ASR/PSNR, while robustness cells report recovered ASR/PSNR.

Variant	Prot.	Gaussian	JPEG	Crop	Resize
Full	81.82/28.94	19.25/35.24	12.97/41.60	22.94/59.51	27.41/29.35
w/o alloc.	77.66/29.57	16.17/35.69	4.58/42.94	22.02/59.74	26.49/29.49
w/o sync.	81.82/28.94	19.25/35.24	12.97/41.60	77.51/29.09	27.19/29.38
w/o state ID	81.82/28.94	22.98/34.96	15.20/41.37	30.03/56.78	27.36/29.39
w/o DA rec.	81.82/28.94	22.84/34.98	14.11/41.39	61.49/29.10	27.13/29.38
w/o key map	81.82/28.94	78.60/27.19	81.49/27.25	76.69/28.67	72.56/28.12

since removing it reduces protection ASR from 81.82% to 77.66% while slightly increasing PSNR; block-phase synchronization, state identification, and distortion-aware recovery mainly support cropping robustness, where recovery degrades from 22.94/59.51 to 77.51/29.09, 30.03/56.78, and 61.49/29.10 when each module is removed; and keyed mapping is essential for preserving the perturbation–recovery correspondence, because disabling it leaves high recovered ASR under Gaussian, JPEG, Crop, and Resize distortions (78.60%, 81.49%, 76.69%, and 72.56%). Overall, robust reversibility arises from the joint effect of keyed coupling, adaptive allocation, and distortion-aware recovery rather than from a single heuristic; full metrics, including SSIM and file expansion rate, are provided in Appendix Tables 16 and 17.

5 Conclusion

This paper presents SRAP-JPEG, a distortion-aware JPEG domain framework for robust reversible adversarial protection of JPEG images over lossy channels. SRAP-JPEG targets at several contradictory objectives, i.e., the protected JPEG images should remain adversarially effective, perceptual quality, compact file-size, key-controlled recoverable and robust against channel attacks for lossy channels. It implements directly on quantized DCT coefficients by decoupling Y-Channel adversarial perturbations from key-controlled Cb/Cr recovery records, and exploits coefficient-adaptive allocation for distortion-aware recovery. Experiments on ImageNet and MS-COCO show that the proposed SRAP-JPEG could achieve superior performance in terms of ASR against classification and vision-language retrieval, exact recovery for lossless channels, and high-fidelity recovery for lossy channels. Ablation study validates the effectiveness of the involved components for robust recovery. And the proposed robust reversible adversarial embedding mechanism could find applications in dataset and copyright protection of JPEG images for both lossless and lossy channels.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China under Grant U22A2030 and Grant U23B2022, and in part by the Guangdong Major Project of Basic and Applied Basic Research under Grant 2023B0303000010.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[3] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[4] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

[5] Dahao Fu, Xiaoyi Zhou, Liaoran Xu, Kaiyue Hou, and Xianyi Chen. Robust reversible watermarking by fractional order zernike moments and pseudo-zernike moments. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7310–7326, 2023.

[6] Jiayang Liu, Weiming Zhang, Kazuto Fukuchi, Youhei Akimoto, and Jun Sakuma. Unauthorized ai cannot recognize me: reversible adversarial example. *arXiv preprint arXiv:1811.00189*, 2018.

[7] Lizhi Xiong, Yue Wu, Peipeng Yu, and Yuhui Zheng. A black-box reversible adversarial example for authorizable recognition to shared images. *Pattern Recognition*, 140:109549, 2023.

[8] Jiajie Zhu, Xia Du, Jizhe Zhou, Chi-Man Pun, Qizhen Xu, and Xiaoyuan Liu. Dp-rae: A dual-phase merging reversible adversarial example for image privacy protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 671–680, 2024.

[9] Fan Xing, Zhuo Tian, Xuefeng Fan, and Xiaoyi Zhou. Raencoder: A label-free reversible adversarial examples encoder for dataset intellectual property protection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20665–20674, 2025.

[10] Jiawei Zhang, Jinwei Wang, Hao Wang, and Xiangyang Luo. Self-recoverable adversarial examples: A new effective protection mechanism in social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):562–574, 2022.

[11] Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, and Talal Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, 1997. doi: 10.1109/83.650120.

[12] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, Rebecca Willett, and Mahdi Soltanolkotabi. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. doi: 10.1109/JSAIT.2020.2991563.

[13] Alejandra Menendez-Ortiz, Claudia Feregrino-Uribe, Rogelio Hasimoto-Beltran, and Jose Juan Garcia-Hernandez. A survey on reversible watermarking for multimedia content: A robustness overview. *IEEE Access*, 7:132662–132681, 2019. doi: 10.1109/ACCESS.2019.2940972.

[14] Ziquan Huang, Bingwen Feng, and Shijun Xiang. Robust reversible image watermarking scheme based on spread spectrum. *Journal of Visual Communication and Image Representation*, 93:103808, 2023. doi: 10.1016/j.jvcir.2023.103808.

[15] Xingyuan Liang and Shijun Xiang. Robust reversible watermarking of JPEG images. *Signal Processing*, 224:109582, 2024. doi: 10.1016/j.sigpro.2024.109582.

[16] Haodong Zhang, Chi Man Pun, and Xia Du. Reversible adversarial examples with beam search attack and grayscale invariance. *arXiv preprint arXiv:2306.11322*, 2023.

[17] Xia Du, Jiajie Zhu, Jizhe Zhou, Chi-man Pun, Zheng Lin, Cong Wu, Zhe Chen, and Jun Luo. Dp-trae: A dual-phase merging transferable reversible adversarial example for image privacy protection. *IEEE Transactions on Dependable and Secure Computing*, 2025.

[18] Zhaoxia Yin, Hua Wang, Li Chen, Jie Wang, and Weiming Zhang. Reversible adversarial attack based on reversible image transformation. *arXiv preprint arXiv:1911.02360*, 2019.

[19] Zhaoxia Yin, Li Chen, Wanli Lyu, and Bin Luo. Reversible attack based on adversarial perturbation and reversible data hiding in yuv colorspace. *Pattern Recognition Letters*, 166:1–7, 2023.

[20] Yucheng Fan, Zhaoxia Yin, Jiawei Chen, and Wanli Lyu. Efficient and transferable reversible adversarial attacks utilizing yuv color space. *Neurocomputing*, page 131088, 2025.

[21] Li Chen, Shaowei Zhu, Abel Andrew, and Zhaoxia Yin. Reversible attack based on local visible adversarial perturbation. *Multimedia Tools and Applications*, 83(4):11215–11227, 2024.

[22] Yiyi Xie, Yuqian Zhou, Tao Wang, Wenying Wen, Shuang Yi, and Yushu Zhang. Reversible gender privacy enhancement via adversarial perturbations. *Neural Networks*, 172:106130, 2024.

[23] Zhen Chen, Xiuli Chai, Zhihua Gan, Binjie Wang, and Yushu Zhang. Rae-vwp: a reversible adversarial example-based privacy and copyright protection method of medical images for internet of medical things. *IEEE Internet of Things Journal*, 2024.

[24] Zeyu Zhao, Ke Xu, Tanfeng Sun, and Xinghao Jiang. Srap: Robust and transferable self-reversible adversarial patch for image privacy protection. *IEEE Transactions on Information Forensics and Security*, 21:3689–3702, 2026.

[25] Mingfu Xue, Yinghao Wu, Yushu Zhang, Jian Wang, and Weiqiang Liu. Dataset authorization control: protect the intellectual property of dataset via reversible feature space adversarial examples. *Applied Intelligence*, 53(6):7298–7309, 2023.

[26] Fan Xing, Xiaoyi Zhou, Hongli Peng, Xuefeng Fan, Wenbao Han, and Yuqing Zhang. Tae-rwp: Traceable adversarial examples with recoverable warping perturbation. *International Journal of Intelligent Systems*, 2024(1):6054172, 2024.

[27] Dahao Fu, Jiangqun Ni, and Jian Zhang. Jpeg-rae: Reversible adversarial example for privacy and copyright protection of jpeg images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9150–9158, 2025.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909, 2025.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[33] Andrew Howard, Andrey Zhmoginov, Liang-Chieh Chen, Mark Sandler, and Menglong Zhu. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In *Proc. cvpr*, pages 4510–4520. IEEE, 2018.

[34] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[37] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[38] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[40] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[41] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018.

A Detailed JPEG-Domain Adversarial Optimization

Adversarial objectives. The adversarial objective depends on the target surrogate model. For conventional image classification, let $l_k(\mathbf{x})$ denote the logit vector produced by the k-th surrogate classifier, w_k denote its fusion weight, and L denote the number of surrogate classifiers. The fused prediction is

$$\mathbf{z}(\mathbf{x}) = \sum_{k=1}^{L} w_k \, l_k(\mathbf{x}), \tag{10}$$

and, given the ground-truth label y, the classification-oriented adversarial loss is

$$L_{\mathrm{adv}} = \mathrm{CE}(\mathbf{z}(\mathbf{x}), y)\,. \tag{11}$$

For vision-language models such as CLIP, let $f_I(\mathbf{x}) \in \mathbb{R}^d$ and $f_T(\mathbf{t}) \in \mathbb{R}^d$ denote image and text embeddings. The semantic adversarial loss is

$$L_{\mathrm{adv}} = -\frac{f_I(\mathbf{x})^\top f_T(\mathbf{t})}{\|f_I(\mathbf{x})\|_2 \, \|f_T(\mathbf{t})\|_2}. \tag{12}$$

Maximizing this objective reduces the semantic alignment between the protected image and its paired text. For simplicity, L_{adv} denotes either adversarial objective in the following derivation.

Gradient transition. The image representation follows the JPEG processing pipeline $\mathbf{X} \to \mathbf{X}_{\mathrm{YCbCr}} \to \mathbf{X}_{\mathrm{DCT}} \to \mathbf{X}_{\mathrm{Quant}} \to \mathbf{X}_{\mathrm{JPEG}}$. Since the adversarial objective is computed in the RGB domain while optimization is performed in the JPEG coefficient domain, gradients must be propagated through this transformation chain. By the chain rule, the gradient with respect to the JPEG representation is

$$\frac{\partial L_{\mathrm{adv}}}{\partial \mathbf{X}_{\mathrm{JPEG}}} = \frac{\partial L_{\mathrm{adv}}}{\partial \mathbf{X}} \cdot \frac{\partial \mathbf{X}}{\partial \mathbf{X}_{\mathrm{YCbCr}}} \cdot \frac{\partial \mathbf{X}_{\mathrm{YCbCr}}}{\partial \mathbf{X}_{\mathrm{DCT}}} \cdot \frac{\partial \mathbf{X}_{\mathrm{DCT}}}{\partial \mathbf{X}_{\mathrm{Quant}}}. \tag{13}$$

The RGB-to-YCbCr conversion and block-wise DCT are linear transforms, so their gradients can be computed analytically. In standard JPEG compression, rounding in quantization is non-differentiable. SRAP-JPEG avoids differentiating through this rounding by optimizing the quantized coefficient representation directly; the remaining conversion between quantized coefficients and DCT-domain values is multiplication or division by the quantization table. Appendix B gives the detailed differentiability argument.

JPEG-domain update and reversible recording. Perturbations are restricted to valid Y-channel locations determined by the keyed cross-channel mapping and coefficient-adaptive capacity allocation. Let ${\mathbf{x}_{\mathrm{JPEG}}}^{(t)} = \{{\mathbf{x}^{Y}_{\mathrm{JPEG}}}^{(t)}, {\mathbf{x}^{Cb}_{\mathrm{JPEG}}}^{(t)}, {\mathbf{x}^{Cr}_{\mathrm{JPEG}}}^{(t)}\}$ denote the JPEG coefficients at iteration t, and let $\delta_Y^{(t)}$ denote the perturbation applied to the Y-channel JPEG coefficients. The signed-gradient update is

$$\delta_Y^{(t+1)} = \delta_Y^{(t)} + \mathbf{M}^Y \odot \mathrm{sign}\left(\frac{\partial L_{\mathrm{adv}}}{\partial {\mathbf{X}^{Y}_{\mathrm{JPEG}}}^{(t)}}\right), \tag{14}$$

where $\mathbf{M}^Y$ is the binary mask indicating valid perturbation locations. To enforce bounded perturbations, we project δ_Y into the feasible range:

$$\delta_Y^{(t+1)} = \mathrm{clip}\Big(\delta_Y^{(t+1)},\, -\alpha,\, \alpha\Big)\,. \tag{15}$$

The updated Y-channel JPEG coefficients are then obtained by

$${\mathbf{X}^{Y}_{\mathrm{JPEG}}}^{(t+1)} = {\mathbf{X}^{Y}_{\mathrm{JPEG}}}^{(0)} + \delta_Y^{(t+1)}. \tag{16}$$

Meanwhile, each Y-channel perturbation is synchronously mapped to its paired chrominance coefficient via the secret-key mapping π_K. For each valid pair (p, q) with $q = \pi_K(p)$, the reversible record is

$$\delta_{Cb/Cr}^{(t+1)}(q) = \delta_Y^{(t+1)}(p), \tag{17}$$

which is embedded into the reserved interval of the Cb/Cr channel. Accordingly, the chrominance coefficients are updated as

$$\mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr\,(t+1)} = \mathbf{X}_{\mathrm{JPEG}}^{Cb/Cr\,(0)} + \delta_{Cb/Cr}^{(t+1)}. \tag{18}$$

After T iterations, the final protected JPEG representation is

$$\mathbf{X}_{\mathrm{JPEG}}^{\mathrm{adv}} = \left\{ \mathbf{X}_{\mathrm{JPEG}}^{Y}{}^{(T)}, \mathbf{X}_{\mathrm{JPEG}}^{Cb}{}^{(T)}, \mathbf{X}_{\mathrm{JPEG}}^{Cr}{}^{(T)} \right\}, \tag{19}$$

which is then encoded into the protected JPEG image.

B Differentiability of JPEG-Domain Gradient Propagation

Proposition. The coefficient-gradient computation used by SRAP-JPEG is differentiable with respect to the optimized quantized DCT coefficients, excluding discrete update operations that are outside the JPEG transformation chain. This follows because the gradient path contains only color conversion, block-wise DCT/IDCT, and quantization-table scaling, while the non-differentiable rounding step of standard JPEG quantization is not invoked during the SRAP-JPEG coefficient update.

Color conversion. For each pixel, the RGB-to-YCbCr transform can be written as the affine map

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.168736 & -0.331264 & 0.5 \\ 0.5 & -0.418688 & -0.081312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix}. \tag{20}$$

Its Jacobian is the constant conversion matrix above, and the constant chrominance offset has zero derivative. Therefore, both RGB-to-YCbCr conversion and its inverse are differentiable affine transforms.

Block-wise DCT. For an 8×8 block $\mathbf{B}$, the two-dimensional DCT coefficient at frequency (u, v) is

$$D_{u,v} = \frac{1}{4} c(u) c(v) \sum_{i=0}^{7} \sum_{j=0}^{7} B_{i,j} \cos\left(\frac{(2i+1)u\pi}{16} \right) \cos\left(\frac{(2j+1)v\pi}{16} \right), \quad u, v \in \{0, \ldots, 7\}, \tag{21}$$

where $c(0) = 1/\sqrt{2}$ and $c(k) = 1$ for $k > 0$. Each coefficient is a fixed weighted sum of the block pixels, so the DCT is a linear transform with a constant Jacobian. The inverse DCT is also linear and differentiable.

Quantization-table scaling. The non-differentiability of ordinary JPEG compression comes from the rounding step in quantization:

$$Z_{u,v} = \mathrm{round}\left(\frac{D_{u,v}}{Q_{u,v}} \right), \qquad \hat{D}_{u,v} = Z_{u,v} Q_{u,v}, \tag{22}$$

where $\mathbf{Q}$ is the JPEG quantization table. Rounding discards the fractional remainder of $D_{u,v}/Q_{u,v}$ and is therefore not differentiable.

SRAP-JPEG does not differentiate through this standard rounding operation when computing coefficient gradients. Instead, the optimized variable is the quantized coefficient $Z_{u,v}$ itself. The DCT-domain value used for reconstruction is obtained by dequantization,

$$\hat{D}_{u,v} = Z_{u,v} Q_{u,v}, \qquad \frac{\partial \hat{D}_{u,v}}{\partial Z_{u,v}} = Q_{u,v}. \tag{23}$$

Equivalently, when a DCT-domain value lies on the JPEG quantization lattice, $\hat{D}_{u,v} = Z_{u,v} Q_{u,v}$, dividing by the quantization table recovers $Z_{u,v}$ exactly and produces no fractional remainder:

$$\frac{\hat{D}_{u,v}}{Q_{u,v}} = Z_{u,v}, \qquad \frac{\partial}{\partial \hat{D}_{u,v}} \left(\frac{\hat{D}_{u,v}}{Q_{u,v}} \right) = \frac{1}{Q_{u,v}}. \tag{24}$$

Thus, the SRAP-JPEG gradient path contains only multiplication or division by nonzero quantization-table entries. The remainder-discarding operation that makes standard JPEG quantization non-differentiable is absent from this path.

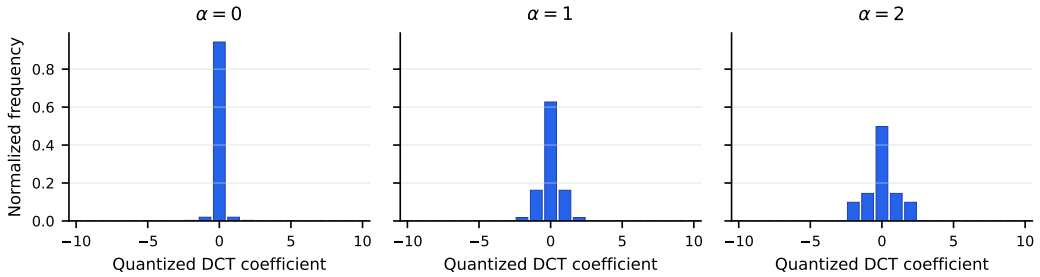

Figure 7: Representative histograms of low-frequency chrominance quantized DCT coefficients under embedding strengths $\alpha = 0$, $\alpha = 1$, and $\alpha = 2$. The three states exhibit different near-zero concentration patterns, which makes deterministic state identification possible without storing extra side information.

Entropy coding. Entropy coding converts the final quantized coefficients into a compressed bitstream using lossless symbol coding. It does not alter the numerical coefficient values used to evaluate L_{adv} and does not participate in the chain-rule computation of coefficient gradients. Consequently, it is outside the differentiable optimization path considered by SRAP-JPEG.

Combining the affine YCbCr transform, the linear DCT/IDCT transform, and the differentiable quantization-table scaling shows that the coefficient-gradient computation in Eq. (13) is differentiable with respect to the optimized SRAP-JPEG coefficient variables.

C Embedding-State Identification

Motivation. SRAP-JPEG does not store the selected embedding strength α as explicit side information in the protected bitstream. Instead, the decoder estimates $\hat{\alpha} \in \{0, 1, 2\}$ from the synchronized chrominance JPEG coefficients before perturbation extraction. This design keeps the recovery pipeline self-contained: after block-phase synchronization, the decoder only needs the shared key and the coefficient statistics of the received image to identify whether no record is present, whether records use strength 1, or whether records use strength 2.

Histogram separability. The state can be identified from the near-zero histogram of low-frequency Cb/Cr quantized DCT coefficients. During embedding, original nonzero chrominance coefficients are shifted outside $[-\alpha, \alpha]$, while reversible records are placed inside this reserved interval. Consequently, different states produce different concentrations around zero: the unembedded state mainly contains the natural JPEG zero peak, the $\alpha = 1$ state increases the mass at $\{-1, 0, 1\}$, and the $\alpha = 2$ state further changes the mass at $\{-2, -1, 0, 1, 2\}$. Fig. 7 shows representative histograms for the three states, illustrating that the embedding strength is visible from simple quantized-DCT statistics.

Deterministic template rule. We implement the analyzer as a deterministic histogram-template classifier. Let Ω_C denote the synchronized low-frequency Cb/Cr candidate coefficient set used for state analysis, and let $\bar{h}(v)$ be the normalized histogram over integer bins $v \in [-10, 10]$:

$$\bar{h}(v) = \frac{1}{|\Omega_C|} \sum_{c \in \Omega_C} \mathbf{1}[\mathrm{clip}(c, -10, 10) = v] \,. \tag{25}$$

The analyzer uses only near-zero histogram features,

$$\phi(\tilde{\mathbf{x}}_{\mathrm{JPEG}}) = \left[\bar{h}(-2), \bar{h}(-1), \bar{h}(0), \bar{h}(1), \bar{h}(2), \sum_{|v| \le 1} \bar{h}(v), \sum_{|v| \le 2} \bar{h}(v) \right] , \tag{26}$$

and selects the closest state template:

$$\hat{\alpha} = \arg \min_{a \in \{0,1,2\}} \left\| \phi(\tilde{\mathbf{x}}_{\mathrm{JPEG}}) - \mu_a \right\|_2 , \tag{27}$$

where μ_a is the fixed reference template obtained by averaging the same histogram feature over reference samples generated with embedding strength a. The same idea can also be implemented with hand-tuned thresholds on the near-zero mass or with a lightweight classifier trained on histogram features. In this paper, we use the deterministic template rule above because it is simple, reproducible, and does not introduce an additional learned recovery model.

D Detailed Experimental Setup

Comparison methods. We compare SRAP-JPEG with five representative reversible adversarial protection baselines: BIM-RAE [6], B-RAE [7], GQBE-RAE [8], SRAE [10], and JPEG-RAE [27]. These methods cover iterative optimization-based, black-box, generative, and encoder-based frameworks. For a fair comparison, all competing methods are evaluated under the same data, model, and metric settings whenever applicable. We follow the official implementations or the configurations reported in the original papers, and only make minor adjustments when necessary to maintain comparable perceptual distortion levels.

Baseline configurations. For BIM-RAE, we use 2×2 superpixels, set the perturbation budget to $\epsilon = 8$, and perform 10 iterations. For B-RAE, we set the perturbation budget to $\epsilon = 6$. For GQBE-RAE, we adopt $\epsilon = 3/255$. For SRAE and JPEG-RAE, we use the settings reported in their original papers. Unless otherwise stated, parameters are selected such that protected images from different methods have comparable visual quality, with PSNR controlled at approximately 30 dB.

Datasets and models. We conduct experiments on ILSVRC 2012 [28] and MS-COCO [29], using 1,000 images from each dataset. For ILSVRC 2012, we evaluate transferability against 10 image recognition models: ResNet-34, ResNet-50, and ResNet-152 [31], DenseNet-121 [32], MobileNet-v2 [33], MobileNet-v3 [34], VGG-16 and VGG-19 [35], AlexNet [36], and Inception-v3 [37]. For MS-COCO, we evaluate multimodal protection using CLIP-based image-text retrieval models [2]; each image is paired with its associated captions for cross-modal evaluation.

Preprocessing and SRAP-JPEG settings. Unless otherwise stated, all input images are resized to 224×224 and compressed with JPEG quality factor Q=75; the controlled experiments use 4:4:4 JPEG. The default SRAP-JPEG configuration uses batch size 1, threshold $\tau = 4000$, and 10 optimization iterations.

Candidate positions and keyed correspondence. Let $S_m(\mathbf{Q})$ denote the set of zero-based block-DCT coordinates $(u, v) \in \{0, \ldots, 7\}^2$ corresponding to the m smallest entries of an 8×8 quantization table $\mathbf{Q}$. Using the standard JPEG luminance and chrominance quantization tables, $\mathbf{Q}_Y^{\mathrm{std}}$ and $\mathbf{Q}_C^{\mathrm{std}}$, we define

$$\begin{aligned} \mathcal{P}_Y &= S_{26}(\mathbf{Q}_Y^{\mathrm{std}}), \\ \mathcal{P}_C &= \big(\{\mathrm{Cb}\} \times S_{13}(\mathbf{Q}_C^{\mathrm{std}})\big) \\ &\qquad \cup \big(\{\mathrm{Cr}\} \times S_{13}(\mathbf{Q}_C^{\mathrm{std}})\big). \end{aligned} \tag{28}$$

Thus, there are 26 Y-channel positions and 26 chrominance slots, comprising 13 positions in each of Cb and Cr. The channel tags distinguish slots with identical block-DCT coordinates in different chrominance channels. Both candidate sets are fixed and public; the secret key specifies an arbitrary bijection $\pi_K : \mathcal{P}_Y \to \mathcal{P}_C$ within spatially corresponding blocks. In our experiments, one such bijection was randomly sampled and kept fixed across the evaluated images. Smaller standard quantization values correspond to finer quantization and motivate the selection of these coefficient positions. The fixed candidate sets are distinct from the image-dependent validity mask in Sec. 3.3, which retains only pairs with a nonzero Y coefficient and a zero chrominance coefficient.

Compute resources. All SRAP-JPEG experiments were run on an NVIDIA A100 GPU with 80 GB memory and CUDA 12.2. SRAP-JPEG processes images with batch size 1. Under this non-parallel setting, without image-level parallelism, the average protection time is 1.38427 seconds per image. Thus, generating protected images for 1,000 images takes approximately 23.1 minutes on a single A100 GPU, excluding baseline evaluation and additional robustness sweeps. The average recovery time is 0.003818 seconds per image. These runtimes are conservative single-image estimates; using image-level parallelism or multiple GPUs can further reduce the wall-clock time.

Table 5: ImageNet classification attack success rate (ASR%) on CNN backbones.

Method	RN-34	RN-50	DN-121	Mob-v2	Mob-v3	VGG-16	AlexNet	RN-152	Inc-v3	VGG-19
BIM-RAE	78.2	73.7	80.1	71.8	81.1	76.2	47.6	76.4	74.7	77.5
B-RAE	85.7	81.4	72.6	91.7	89.9	93.7	71.1	80.2	79.4	92.4
GQBE-RAE	85.2	83.0	88.6	79.7	89.9	82.5	60.4	86.2	88.3	83.2
SRAE	48.8	99.0	32.8	88.7	33.8	85.2	40.2	43.2	21.2	94.1
JPEG-RAE	86.4	83.9	90.1	80.6	90.0	82.8	59.4	89.6	92.6	84.2
SRAP-JPEG	85.7	84.6	90.0	78.6	90.3	80.7	49.2	88.5	88.0	82.6

Table 6: ImageNet classification attack success rate (ASR%) on Transformer-style and hybrid vision backbones.

Method	ViT	PiT	Swin	Twins	CoaT	DeiT	PvT	LeViT	Visformer	MobileViT
BIM-RAE	81.9	58.4	59.2	61.3	64.4	56.5	66.8	59.4	66.1	76.9
B-RAE	27.6	26.4	36.0	41.0	58.5	19.5	80.5	62.0	41.8	78.3
GQBE-RAE	82.3	63.7	62.6	70.2	75.4	60.6	79.9	71.9	75.2	86.0
SRAE	12.3	4.8	14.1	14.1	20.4	4.1	11.0	14.5	11.1	9.1
JPEG-RAE	86.6	76.1	77.1	80.5	82.2	72.2	85.7	80.9	81.2	85.7
SRAP-JPEG	66.0	78.0	73.0	83.7	62.3	79.6	80.1	86.1	63.1	83.3

Evaluation metrics. For ImageNet-based classification, Attack Success Rate (ASR) measures the proportion of images that successfully fool the target models. For perceptual and recovery quality, we report Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [38, 39] for protected and recovered images. File expansion rate measures JPEG-domain compatibility as the ratio between the protected JPEG file size and the original file size. For COCO-based multimodal retrieval, we use text-retrieval and image-retrieval recall metrics widely adopted in cross-modal retrieval [40, 41]. TR@1, TR@5, and TR@10 denote recall when retrieving the ground-truth text given an image query, while IR@1, IR@5, and IR@10 denote recall when retrieving the ground-truth image given a text query.

E Detailed Attack Ability Results

This section provides the complete attack-ability results summarized in the main paper. Table 5 reports transferability on 10 CNN-based ImageNet classifiers, and Table 6 further evaluates 10 Transformer-style or hybrid vision backbones. Table 7 reports CLIP-based text-to-image and image-to-text retrieval on MS-COCO, where lower recall indicates stronger disruption of cross-modal semantic matching.

On ImageNet, SRAP-JPEG remains competitive with the strongest reversible baselines across both CNN and Transformer-style backbones. On the 10 CNN classifiers, SRAP-JPEG achieves an average ASR of 81.8%, close to B-RAE and JPEG-RAE and substantially higher than BIM-RAE and SRAE. On the 10 Transformer-style or hybrid backbones, SRAP-JPEG obtains a 75.5% average ASR, outperforming BIM-RAE, B-RAE, GQBE-RAE, and SRAE, while remaining below JPEG-RAE. The per-model results show that SRAP-JPEG transfers well to several non-CNN architectures, including PiT, Twins, DeiT, PvT, LeViT, and MobileViT. These results indicate that SRAP-JPEG preserves broad attack transferability despite stricter constraints: perturbations must be embedded in quantized JPEG coefficients, paired with reversible records, and kept compatible with later distortion-aware recovery.

On MS-COCO, SRAP-JPEG consistently reduces retrieval recall across all three OpenCLIP backbones. Compared with AnyAttack, SRAP-JPEG lowers R@Mean from 42.1 to 17.8 on ViT-B/32, from 47.0 to 19.8 on ViT-B/16, and from 52.0 to 39.3 on ViT-L/14. The average R@Mean over the three backbones is reduced from 47.0 to 25.6, indicating that the protected JPEG images disrupt both text retrieval and image retrieval. Although ViT-L/14 is more resistant than the two smaller backbones, SRAP-JPEG still produces a clear recall drop while preserving the reversible-protection requirements considered in this paper.

Table 7: COCO multimodal retrieval recall. Lower recall indicates stronger attack performance.

Model	Attack Method	Text Retrieval				Image Retrieval				Overall
		TR@1	TR@5	TR@10	TR@Mean	IR@1	IR@5	IR@10	IR@Mean	R@Mean
ViT-B/32	Clean	73.1	94.1	97.3	88.2	57.8	86.8	94.0	79.5	83.9
	AnyAttack	27.0	45.8	53.5	42.1	23.8	46.1	56.2	42.0	42.1
	SRAP-JPEG	9.9	20.0	27.2	19.0	6.0	17.6	25.9	16.5	17.8
ViT-B/16	Clean	73.6	93.8	97.7	88.4	58.1	86.6	94.0	79.6	84.0
	AnyAttack	30.1	51.1	59.7	47.0	28.3	51.0	61.9	47.1	47.0
	SRAP-JPEG	12.4	24.4	33.5	23.4	8.7	23.8	34.4	16.1	19.8
ViT-L/14	Clean	70.9	92.8	97.7	87.1	53.0	83.1	91.7	75.9	81.5
	AnyAttack	37.5	51.0	68.9	52.5	30.3	55.8	68.0	51.4	52.0
	SRAP-JPEG	24.4	45.4	55.5	41.8	16.9	39.8	53.8	36.8	39.3

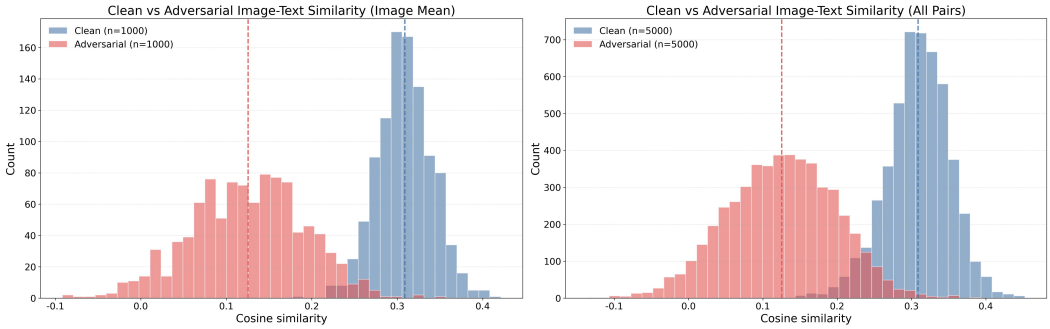

Figure 8: Comparison of cosine similarity between images and their corresponding texts for clean and protected samples. SRAP-JPEG reduces image-text alignment in the shared embedding space, supporting its attack effect on CLIP-based retrieval models.

F Additional Attack Ability Analysis

Image-text feature alignment. Fig. 8 compares the cosine similarity between image features and their paired text features for clean and protected samples. Clean image-text pairs preserve high similarity in the shared CLIP embedding space, while SRAP-JPEG protected images show a clear similarity decrease. This feature-level evidence supports the COCO retrieval degradation reported in Table 7.

G Additional Visual Quality Analysis

Fig. 9 provides a qualitative comparison using a representative image of a small dog held in a hand. The protected images produced by BIM-RAE and B-RAE exhibit relatively mild visual distortion, while SRAE and GQBE introduce more noticeable structural or texture-like artifacts. JPEG-RAE and SRAP-JPEG both preserve the overall semantic appearance of the image well, and the protected sample generated by SRAP-JPEG remains visually natural despite operating under additional robustness and security constraints. The magnified difference maps further show that SRAP-JPEG distributes perturbations in a structured manner rather than producing severe global corruption. Moreover, the recovered images of SRAP-JPEG are visually indistinguishable from the cover image, which is consistent with the quantitative recovery results in Table 3.

H Poisoning Attack

Beyond misleading pretrained models at inference time, SRAP-JPEG can also protect image data from unauthorized model training by acting as a poisoning source. To verify this property, we construct three training sets for ResNet-50 under a controlled setting: clean JPEG images, SRAP-

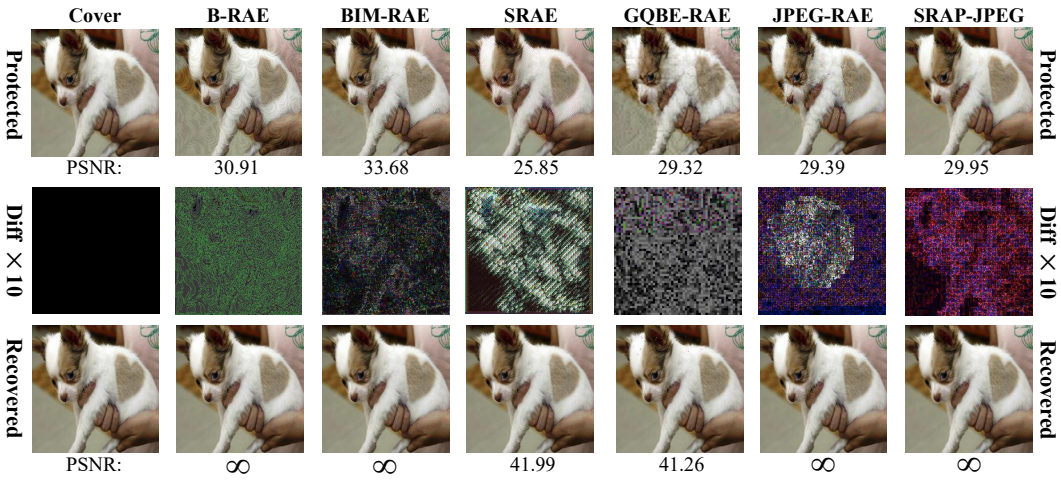

Figure 9: Qualitative visual quality comparison of different RAE schemes. From top to bottom: protected images, magnified difference maps ($\times 10$), and recovered images. The example shows that SRAP-JPEG preserves the overall appearance of the cover image while maintaining exact recovery, and its perturbation pattern is more structured and visually controllable than methods that introduce stronger visible artifacts.

JPEG protected images, and random-noise-corrupted images with matched perturbation intensity. Each set contains 10,000 images from 10 classes, with 1,000 images per class. The clean and noise groups serve as references for normal training and perturbation-magnitude control, respectively, whereas the SRAP-JPEG group is used to test whether structured reversible adversarial protection can degrade downstream learning.

As shown in Fig. 10, all three models fit their training sets well, but their optimization and generalization behaviors differ markedly. After 20 epochs, the training loss reaches 0.0087, 0.0211, and 0.0069 for the clean, SRAP-JPEG, and noise groups, respectively, while the corresponding training accuracies are 99.95%, 99.87%, and 99.98%. Although the model trained on SRAP-JPEG protected data still attains high training accuracy, it converges more slowly and to a noticeably higher training loss than the clean and matched-noise baselines. The gap becomes more evident on the validation set. In the later epochs, the validation loss of the SRAP-JPEG group remains around 0.39–0.42, whereas the clean and noise groups stabilize near 0.23–0.28. A similar trend is observed in validation accuracy: after 20 epochs, the clean, SRAP-JPEG, and noise groups achieve 92.34%, 87.86%, and 93.30%, respectively. Therefore, training on SRAP-JPEG protected data reduces validation accuracy by 4.48 percentage points relative to clean training and by 5.44 percentage points relative to the matched-noise baseline. These results demonstrate that SRAP-JPEG does not merely add visually comparable perturbation energy; instead, it alters the training signal in a structured adversarial manner and significantly weakens the generalization ability of the resulting model.

I Additional Expansion Rate Analysis

Table 8 reports the effect of varying the JPEG quality factor from 50 to 90 under the same ImageNet CNN attack setting. As the quality factor increases, both the cover and protected file sizes grow steadily, from 8.81 KB and 11.62 KB at $Q{=}50$ to 23.66 KB and 28.84 KB at $Q{=}90$, while the protected-image quality also improves, with PSNR increasing from 25.19 dB to 32.69 dB and SSIM rising from 0.78 to 0.91. This gain in visual fidelity is accompanied by a clear reduction in adversarial strength, as the average ASR decreases from 87.64% to 67.08%. In our experiments, $Q{=}75$ serves as a practical default operating point because it provides a balanced trade-off among storage cost, visual quality, and attack performance, yielding a 17.05 KB protected size, 28.92 dB PSNR, 0.85 SSIM, and 81.82% ASR.

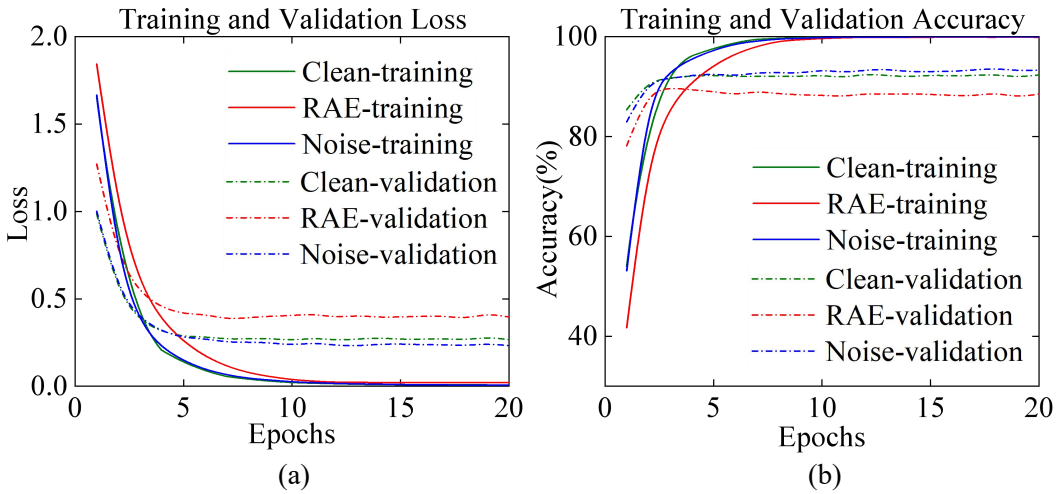

Figure 10: Training and validation curves under clean, SRAP-JPEG-protected, and matched-noise training data. Although all models achieve high training accuracy, the model trained on SRAP-JPEG protected samples converges more slowly and exhibits consistently higher validation loss and lower validation accuracy, demonstrating a clear poisoning effect on downstream learning.

Table 8: Effect of JPEG quality factor on file size, visual quality, and attack success rate of SRAP-JPEG.

QF	50	55	60	65	70	75	80	85	90
Cover (KB)	8.81	9.56	10.28	11.21	12.29	13.42	15.65	18.56	23.66
SRAP (KB)	11.62	12.52	13.39	14.48	15.77	17.05	19.83	23.04	28.84
PSNR (dB)	25.19	25.88	26.52	27.23	28.01	28.92	30.12	31.30	32.69
SSIM	0.78	0.80	0.81	0.83	0.84	0.85	0.87	0.89	0.91
ASR (%)	87.64	86.35	85.36	84.07	83.04	81.82	78.29	75.09	67.08

J Detailed Recovery Robustness Results

Table 9 reports the full visual recovery sweep under Gaussian noise, JPEG recompression, edge cropping, and resizing. SRAP-JPEG achieves lossless recovery without post-processing distortion and consistently provides higher recovered-to-cover PSNR than SRAE after distortion. Under Gaussian noise, SRAP-JPEG recovers to 80.01 dB at noise level 0.5 and remains above SRAE across all noise levels. For protected JPEGs generated at $QF = 75$, exact recovery is observed in the tested aligned recompression settings with $QF \in \{75, 80, 85, 90, 95\}$. This observation is specific to these evaluated conditions and does not establish lossless recovery for arbitrary initial quality factors or recompression histories. The most pronounced gain appears under edge cropping: SRAP-JPEG recovers images with PSNR around 59.4–59.8 dB over all cropping ratios, whereas SRAE mostly stays near 22–23 dB after cropping. Under resizing, SRAP-JPEG remains stronger at most resolutions but approaches the SRAE recovery quality under the most severe downsampling, indicating a practical robustness boundary.

Table 10 further examines whether recovery can compensate for external distortions, rather than only removing the adversarial perturbation. Under Gaussian noise, SRAP-JPEG often produces recovered images with higher PSNR than both the distorted protected image and the directly corrupted cover image, indicating that recovery suppresses part of the external noise as well. SRAE does not show this behavior, as its recovered PSNR is close to or lower than the noisy-protected reference. In the same $QF = 75$ aligned recompression experiment, exact recovery is observed at the tested moderate quality factors, while recovery degrades under stronger recompression.

Diagnostic baseline failure. We further diagnose whether clean reversible recovery directly survives post-processing. For each baseline, the protected image is first generated using its original

Table 9: Visual recovery under post-processing distortions. All entries are PSNR (dB) with respect to the original cover image, except that edge-cropping PSNR is computed only over the spatially aligned uncropped region because removed boundary pixels are not observable after cropping. Protection denotes distorted protected images, Recovery denotes recovered outputs, and Clean-Dist. denotes cover images after the same distortion.

Gaussian noise												
Stage	Method	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Protection	SRAE [10]	24.44	24.42	24.34	24.20	24.01	23.78	23.51	23.22	22.91	22.58	22.24
	SRAP-JPEG	**28.94**	**28.86**	**28.60**	**28.22**	**27.74**	**27.19**	**26.17**	**26.03**	**25.44**	**24.87**	**24.31**
Recovery	SRAE [10]	39.40	38.34	36.41	34.45	32.70	31.19	29.88	28.72	27.70	26.78	25.95
	SRAP-JPEG	∞	**80.01**	**62.34**	**40.83**	**37.52**	**35.24**	**33.45**	**31.98**	**30.64**	**29.44**	**28.24**
Reference	Clean-Dist.	∞	45.39	39.86	36.45	34.00	32.10	30.54	29.22	28.08	27.07	26.18
JPEG recompression												
Stage	Method	100	95	90	85	80	75	70	65	60	55	50
Protection	SRAE [10]	24.44	24.02	23.90	23.79	23.67	23.62	23.57	23.56	23.62	23.76	23.96
	SRAP-JPEG	**28.94**	**28.58**	**27.86**	**27.92**	**29.26**	**28.94**	**27.25**	**26.73**	**25.68**	**24.51**	**25.40**
Recovery	SRAE [10]	39.40	32.46	30.69	29.25	28.28	27.57	27.06	26.59	26.13	25.86	25.71
	SRAP-JPEG	∞	∞	∞	∞	∞	∞	**41.60**	**34.63**	**27.99**	**27.47**	**28.53**
Reference	Clean-Dist.	∞	57.37	46.05	44.82	42.39	∞	41.77	37.38	34.75	32.51	30.47
Edge cropping												
Stage	Method	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Protection	SRAE [10]	24.44	24.29	24.15	24.04	23.95	23.90	23.86	23.82	23.79	23.76	23.71
	SRAP-JPEG	**28.94**	**28.76**	**28.88**	**29.11**	**28.43**	**28.67**	**28.82**	**28.70**	**28.61**	**28.58**	**28.49**
Recovery	SRAE [10]	39.40	23.18	22.66	22.88	34.29	22.75	22.85	22.31	22.24	22.49	22.40
	SRAP-JPEG	∞	**59.43**	**59.44**	**59.51**	**59.50**	**59.51**	**59.75**	**59.75**	**59.67**	**59.67**	**59.79**
Reference	Clean-Dist.	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
Resize												
Stage	Method	272	256	240	224	208	192	176	160	144	128	112
Protection	SRAE [10]	28.90	28.61	28.70	24.44	27.98	27.23	26.75	26.08	25.64	25.21	22.61
	SRAP-JPEG	**29.06**	**28.98**	**28.87**	**28.94**	**28.59**	**28.42**	**28.19**	**27.88**	**27.65**	**27.33**	**27.21**
Recovery	SRAE [10]	28.12	28.90	28.96	39.39	28.86	28.77	28.49	28.02	27.74	27.28	27.26
	SRAP-JPEG	**32.11**	**31.61**	**31.27**	∞	**30.37**	**29.83**	**29.39**	**28.92**	**28.35**	**27.80**	**27.03**
Reference	Clean-Dist.	34.42	33.92	33.43	∞	32.34	31.74	31.14	30.48	29.77	29.03	28.10

configuration, then distorted, and finally decoded using the corresponding original recovery procedure without adding SRAP-JPEG-style synchronization or distortion-specific repair. This experiment is intended to separate clean reversibility from robust reversibility: if the reversible records are altered, shifted, or removed by post-processing, the decoder can no longer reliably remove adversarial perturbations or reconstruct the authorized content.

Table 11 shows that clean reversibility does not imply robust reversibility after post-processing. Although several baselines can recover clean protected images, their recovered ASR remains high and their recovered PSNR drops sharply after Gaussian noise, JPEG recompression, cropping, or resizing. This behavior indicates that the reversible records used by clean-bitstream RAE schemes are easily corrupted or misaligned once the protected image is modified. In contrast, SRAP-JPEG preserves usable recovery records under the same distortions, especially under JPEG recompression and cropping.

Composed image-circulation distortions. The single-distortion sweeps above isolate the effect of each operation. To approximate practical image circulation more closely, we also evaluate composed post-processing pipelines that combine resizing, cropping, noise, and JPEG re-saving. This setting tests whether authorized recovery remains usable when the protected JPEG undergoes multiple common transformations before decoding.

Table 12 further shows the difference between distortion-induced attack degradation and genuine authorized recovery. For SRAE, the protected ASR already drops substantially under several composed pipelines, e.g., 23.24% after resizing plus JPEG recompression and 26.90% after cropping, resizing, and JPEG re-saving. Thus, a relatively low recovered ASR in these cases should not be interpreted as successful recovery; much of the adversarial effect has already been weakened by the

Table 10: Distortion-compensation analysis. All entries are PSNR (dB), except that edge-cropping PSNR is computed only over the spatially aligned uncropped region because removed boundary pixels are not observable after cropping. D-Prot. denotes the distorted protected image, Prot. denotes the undistorted protected image, and Clean-Dist. denotes the cover image after the same distortion.

Gaussian noise												
Measure	Method	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
D-Prot./Prot.	SRAE [10]	∞	45.88	40.03	36.55	34.07	32.15	30.59	29.27	28.12	27.11	26.12
	SRAP-JPEG	∞	42.77	38.76	35.87	33.65	31.87	30.38	29.11	28.00	27.02	26.14
Recovered/Cover	SRAE [10]	39.40	38.34	36.41	34.45	32.70	31.19	29.88	28.72	27.70	26.78	25.95
	SRAP-JPEG	∞	**80.01**	**62.34**	**40.83**	**37.52**	**35.24**	**33.45**	**31.98**	**30.64**	**29.44**	**28.24**
Clean-Dist./Cover	Cover	∞	45.39	39.86	36.45	34.00	32.10	30.54	29.22	28.08	27.07	26.18
JPEG recompression												
Measure	Method	100	95	90	85	80	75	70	65	60	55	50
D-Prot./Prot.	SRAE [10]	∞	34.62	31.97	30.22	28.93	27.80	27.04	26.29	25.66	25.12	24.68
	SRAP-JPEG	∞	52.23	43.41	42.39	40.06	∞	39.37	34.80	32.52	30.28	28.24
Recovered/Cover	SRAE [10]	39.40	32.46	30.69	29.25	28.28	27.57	27.06	26.59	26.13	25.86	25.71
	SRAP-JPEG	∞	∞	∞	∞	∞	∞	**41.60**	**34.63**	**27.99**	**27.47**	**28.53**
Clean-Dist./Cover	Cover	∞	57.37	46.05	44.82	42.39	∞	41.77	37.38	34.75	32.51	30.47
Edge cropping												
Measure	Method	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
D-Prot./Prot.	SRAE [10]	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
	SRAP-JPEG	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
Recovered/Cover	SRAE [10]	39.40	23.18	22.66	22.88	34.29	22.75	22.85	22.31	22.24	22.49	22.40
	SRAP-JPEG	∞	**59.43**	**59.44**	**59.51**	**59.50**	**59.51**	**59.75**	**59.75**	**59.67**	**59.67**	**59.79**
Clean-Dist./Cover	Cover	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞
Resize												
Measure	Method	272	256	240	224	208	192	176	160	144	128	112
D-Prot./Prot.	SRAE [10]	26.99	26.58	26.18	∞	25.31	24.86	24.44	24.00	23.60	23.21	22.92
	SRAP-JPEG	32.70	32.20	31.75	30.66	30.06	29.40	28.79	28.07	27.33	26.62	27.21
Recovered/Cover	SRAE [10]	28.12	29.00	28.96	24.44	28.86	28.77	28.49	28.02	27.74	27.28	27.26
	SRAP-JPEG	**32.11**	**31.61**	**31.27**	∞	**30.37**	**29.83**	**29.39**	**28.92**	**28.35**	**27.80**	**27.03**
Clean-Dist./Cover	Cover	34.42	33.92	33.43	∞	32.34	31.74	31.14	30.48	29.77	29.03	28.10

Table 11: Diagnostic recovery of existing RAE baselines after post-processing. Each distorted protected image is decoded with the original recovery procedure of the corresponding method. Each metric cell reports recovered ASR / recovered-to-cover PSNR; lower ASR and higher PSNR indicate better authorized recovery.

Method	Clean	Gaussian 2.5	JPEG QF=70	Crop 0.25	Resize 176	Diagnostic note
BIM-RAE [6]	0.00/∞	74.69/24.49	76.41/28.56	71.20/31.81	73.99/27.21	RDH record corruption / mismatch
B-RAE [7]	0.00/∞	68.35/24.34	67.81/28.10	79.29/30.86	73.20/26.55	RDH record corruption / mismatch
GQBE-RAE [8]	5.75/46.1	83.56/24.03	83.40/27.10	81.11/29.51	84.52/25.76	Auxiliary record alteration
JPEG-RAE [27]	0.00/∞	81.47/24.14	81.99/26.64	80.11/29.47	75.55/26.40	Coefficient-record corruption
SRAE [10]	6.32/39.40	18.05/31.19	37.52/27.06	70.14/22.75	44.82/28.49	Limited distorted recovery
SRAP-JPEG	0.00/∞	19.25/35.24	12.97/41.60	22.94/59.51	27.41/29.35	Distortion-aware recovery

composed distortions before decoding. This is also reflected by the recovered PSNR, which remains low for SRAE across all composed pipelines. By contrast, SRAP-JPEG maintains higher protected ASR before authorization and then reduces recovered ASR while retaining higher recovered PSNR, showing that the recovery process itself remains effective after practical image circulation.

Real-platform circulation. We evaluated the same set of 50 protected images on X, Facebook, and WeChat. Each image was upsampled to 2048×2048 and uploaded once to each platform, yielding 150 upload-download instances. Protection and authorized recovery were evaluated on the 50 downloaded images from each platform, with PSNR averaged over those images. The same recovery configuration and authorized mapping were used without access to the platforms' processing histories or parameters. All downloaded files had different hashes from their corresponding uploads; the upload-to-download PSNR in Table 13 additionally quantifies changes to image content. The

Table 12: Recovery under composed image-circulation distortions. Each pipeline combines common post-processing operations before authorized recovery. Lower recovered ASR and higher recovered-to-cover PSNR indicate better robust reversibility.

Pipeline	Method	Protected ASR	Recovered ASR ↓	Recovered PSNR ↑
Resize $256 \rightarrow 224$ + JPEG QF=70	SRAE [10]	23.24	35.59	27.53
	SRAP-JPEG	76.14	24.86	30.21
Crop 0.10 + JPEG QF=85	SRAE [10]	52.09	60.90	22.40
	SRAP-JPEG	74.08	22.32	59.43
Gaussian 1.0 + JPEG QF=70	SRAE [10]	61.18	38.00	27.04
	SRAP-JPEG	81.14	13.71	38.73
Crop 0.10 + Resize $256 \rightarrow 224$ + JPEG QF=85	SRAE [10]	26.90	48.22	26.42
	SRAP-JPEG	69.51	28.02	31.51

experiment measures the aggregate effect of each tested platform workflow without assuming a particular internal operation.

Table 13: Real-platform circulation of 50 protected images uploaded at 2048×2048. Protected and recovered PSNR are measured against the cover; upload-to-download PSNR compares the uploaded and downloaded images.

	ASR (%)		PSNR (dB)		
Platform	Protected	Recovered	Prot./cover	Rec./cover	Upload/download
X	77.5	6.8	29.36	41.20	45.96
Facebook	77.0	6.2	29.48	41.04	44.01
WeChat	77.2	6.2	29.45	39.86	41.02

Across these workflows, protected ASR remained between 77.0% and 77.5%, while authorized recovery reduced ASR to 6.2%–6.8% and achieved recovered-to-cover PSNR of 39.86–41.20 dB. These results establish performance for the tested upsampled uploads and platform workflows; they do not establish performance for arbitrary platform processing or native high-resolution inputs.

Effect of upload resolution. A separate experiment used the same 10 protected images across all conditions in Table 14. Each image was resized to 512×512, 1024×1024, and 2048×2048 and uploaded once at each resolution to each of the three platforms, yielding 90 upload-download instances. ASR was evaluated over 10 images and 10 models per condition, and PSNR was averaged over the same 10 images. These results are reported separately from the 50-image experiment above.

Table 14: Effect of upload resolution on real-platform circulation using 10 protected images. The same images are evaluated in all nine conditions; PSNR is measured against the cover.

		ASR (%)		PSNR (dB)	
Upload size	Platform	Protected	Recovered	Prot./cover	Rec./cover
512×512	X	78.0	19.0	29.53	33.94
	Facebook	77.0	18.0	30.21	31.61
	WeChat	79.0	26.0	29.74	33.64
1024×1024	X	78.0	19.0	29.54	33.97
	Facebook	80.0	18.0	29.63	33.78
	WeChat	79.0	15.0	29.63	33.70
2048×2048	X	81.0	3.0	29.03	41.43
	Facebook	80.0	3.0	29.13	41.25
	WeChat	79.0	3.0	29.13	40.20

Averaged across platforms, protected ASR was 78.0%, 79.0%, and 80.0% at the three upload sizes, respectively. Recovered ASR decreased from 21.0% to 17.3% and then to 3.0%, while recovered-

to-cover PSNR increased from 33.06 to 33.82 and then to 40.96 dB. Higher upload resolution was therefore associated with better average recovery in this evaluated set, especially at 2048×2048; this observation does not imply a universal resolution–recovery relationship.

Double-JPEG recompression. To examine a different initial quality factor, we generated protected JPEGs at $QF_1 = 80$ and recompressed them at $QF_2 \in \{80, 75, 70, 65, 60\}$. Under the authorized mapping, recovery used the same configuration throughout and took the processed image without being supplied with either quality factor, the number of compression rounds, or other compression-history parameters. The quality factors specify the evaluation conditions rather than additional decoder inputs. Table 15 reports both protection and recovery metrics.

Table 15: Double-JPEG evaluation with protected images generated at $QF_1 = 80$. All PSNR values are measured against the cover. This experiment uses a different initial quality factor from the $QF = 75$ sweep in Table 9.

	ASR (%)		PSNR (dB)	
$QF_1 \to QF_2$	Protected	Recovered	Protected	Recovered
$80 \to 80$	82.16	2.24	29.24	59.67
$80 \to 75$	81.61	2.18	28.75	59.22
$80 \to 70$	81.92	15.50	27.19	38.88
$80 \to 65$	82.13	29.20	25.87	34.60
$80 \to 60$	79.09	65.07	26.97	28.85

Protected ASR remained between 79.09% and 82.16% across these settings, whereas authorized recovery degraded under stronger recompression. The recovered PSNR values of 59.67 and 59.22 dB at $80 \to 80$ and $80 \to 75$ indicate high-fidelity rather than lossless recovery. At $80 \to 60$, recovered ASR increased to 65.07% and recovered PSNR fell to 28.85 dB. This degradation is consistent with stronger quantization damaging the chrominance recovery records. The recovered PSNR of 38.88 dB at $80 \to 70$ belongs to this experiment and is distinct from the 41.60 dB result at $75 \to 70$ in Table 9.

K Detailed Ablation Study

Table 16 and Table 17 provide the full component ablation results. The protection-side metrics include ASR, PSNR, SSIM, and file expansion rate (Exp.), where higher ASR/PSNR/SSIM and lower Exp. are preferred. The recovery-side metrics include recovered ASR and recovered PSNR under Gaussian noise, JPEG recompression, cropping, and resizing.

The protection-side results show that the coefficient-adaptive allocation strategy mainly improves the attack-quality trade-off rather than simply maximizing all metrics. The full SRAP-JPEG achieves an ASR of 81.82%, whereas removing adaptive allocation reduces ASR to 77.66%, a drop of 4.16 percentage points. Meanwhile, the variant without adaptive allocation obtains slightly higher visual quality, with PSNR increasing from 28.94 dB to 29.57 dB and SSIM increasing from 0.85 to 0.87. This behavior is expected: the adaptive allocation module intentionally allocates coefficient capacity to support stronger adversarial protection while maintaining acceptable perceptual quality and the same file expansion rate.

The recovery-side results further demonstrate that robust reversibility is produced by the coupled recovery design rather than by direct inversion alone. Removing block-phase synchronization barely affects Gaussian noise or JPEG recompression, but it causes a severe failure under cropping: Rec. ASR increases from 22.94% to 77.51%, and Rec. PSNR drops from 59.51 dB to 29.09 dB. Removing state identification also degrades recovery across distortions, especially under cropping, where Rec. ASR rises to 30.03% and Rec. PSNR falls to 56.78 dB. The variant without the distortion-aware recovery pipeline suffers a stronger collapse under cropping, reaching 61.49% Rec. ASR and only 29.10 dB Rec. PSNR. Finally, disabling keyed mapping almost prevents meaningful recovery, with Rec. ASR remaining high under Gaussian, JPEG, Crop, and Resize distortions at 78.60%, 81.49%, 76.69%, and 72.56%, respectively.

Table 16: Component ablation of SRAP-JPEG under clean protection, Gaussian noise, and JPEG recompression. Gaussian and JPEG denote recovery after Gaussian noise level 2.5 and JPEG recompression with QF=70, respectively. The variant w/o keyed mapping disables the secret-key-controlled correspondence between perturbation and recovery positions.

Variant	Protection				Gaussian		JPEG	
	ASR ↑	PSNR ↑	SSIM ↑	Exp. ↓	Rec. ASR ↓	Rec. PSNR ↑	Rec. ASR ↓	Rec. PSNR ↑
Full SRAP-JPEG	81.82	28.94	0.85	127	19.25	35.24	12.97	41.60
w/o adaptive allocation	77.66	29.57	0.87	127	16.17	35.69	4.58	42.94
w/o block-phase sync.	81.82	28.94	0.85	127	19.25	35.24	12.97	41.60
w/o state identification	81.82	28.94	0.85	127	22.98	34.96	15.20	41.37
w/o distortion-aware recovery	81.82	28.94	0.85	127	22.84	34.98	14.11	41.39
w/o keyed mapping	81.82	28.94	0.85	127	78.60	27.19	81.49	27.25

Table 17: Component ablation of SRAP-JPEG under cropping and resizing. Crop and Resize denote recovery after crop ratio 0.25 and resize size 176, respectively.

Variant	Protection				Crop		Resize	
	ASR ↑	PSNR ↑	SSIM ↑	Exp. ↓	Rec. ASR ↓	Rec. PSNR ↑	Rec. ASR ↓	Rec. PSNR ↑
Full SRAP-JPEG	81.82	28.94	0.85	127	22.94	59.51	27.41	29.35
w/o adaptive allocation	77.66	29.57	0.87	127	22.02	59.74	26.49	29.49
w/o block-phase sync.	81.82	28.94	0.85	127	77.51	29.09	27.19	29.38
w/o state identification	81.82	28.94	0.85	127	30.03	56.78	27.36	29.39
w/o distortion-aware recovery	81.82	28.94	0.85	127	61.49	29.10	27.13	29.38
w/o keyed mapping	81.82	28.94	0.85	127	76.69	28.67	72.56	28.12

L Security Analysis

Threat model. We consider an unauthorized user who can access the protected JPEG image and knows the SRAP-JPEG algorithm, including the coefficient-selection rule, the recovery pipeline, and the possible embedding strengths, but does not know the secret key K used to instantiate the cross-channel mapping π_K. Such an adversary may attempt to recover the image with a wrong key, replace the keyed correspondence with a random mapping, or first apply common post-processing operations such as JPEG recompression and cropping before recovery. In contrast, an authorized user holds the correct key and can reproduce the perturbation–record correspondence during decoding.

Key-gated recovery evidence. Table 18 summarizes the security-oriented recovery checks used to isolate the role of the secret key. Panel (a) evaluates wrong-key recovery on 1,000 ImageNet images under the default $Q{=}75$ setting and two representative circulation distortions. Panel (b) further tests partial-key sensitivity by preserving only a fraction ρ of the correct perturbation–record pairs and randomly permuting the remaining pairs. Compared with correct-key recovery, wrong-key and random-mapping recovery retain much higher recovered ASR and lower recovered PSNR, indicating that the perturbation–record correspondence cannot be reliably inferred without authorization. The partial-key results further show that recovery improves as more correct pairs are available at the decoder: recovered ASR decreases and recovered PSNR increases with larger ρ. These results support the role of the secret-key-controlled mapping as the authorization boundary for recovery.

Scope and limitation. This analysis supports key-gated recovery rather than cryptographic secrecy. SRAP-JPEG does not encrypt the protected image, and an unauthorized user may still view a visually natural but adversarially protected JPEG. The security role of the key is to control authorized recovery: without the correct perturbation–record correspondence, the decoder cannot reliably remove the adversarial effect or reconstruct high-fidelity cover content, especially after practical image circulation operations that already disturb JPEG coefficients and block alignment.

M Limitations

The secret key in SRAP-JPEG controls the perturbation–record correspondence required for authorized recovery, but it does not provide cryptographic secrecy. An unauthorized user can still view the protected JPEG, and the security goal is to prevent reliable removal of the adversarial effect and

Table 18: Security-oriented recovery analysis. Panel (a) compares authorized recovery with wrong-key and random-mapping recovery. Panel (b) measures sensitivity to the fraction ρ of correct key pairs retained at the decoder. Metric columns report recovered ASR (%; lower is better) and recovered PSNR (dB; higher is better). Randomized settings are averaged over the listed seeds.

(a) Wrong-key recovery

Decoder setting	Clean		JPEG QF=70		Crop ratio=0.25	
	Rec. ASR ↓	Rec. PSNR ↑	Rec. ASR ↓	Rec. PSNR ↑	Rec. ASR ↓	Rec. PSNR ↑
Correct key	0.00	∞	12.97	41.60	22.94	59.51
Wrong key (avg., 5 seeds)	56.52	29.01	61.45	29.17	56.19	28.90
Random mapping	55.78	29.47	61.66	29.16	55.18	28.89

(b) Partial-key sensitivity

Correct-pair ratio ρ	Clean		JPEG QF=70		Notes	
	Rec. ASR ↓	Rec. PSNR ↑	Rec. ASR ↓	Rec. PSNR ↑	Seeds	Decoder pairs
0%	57.07	28.80	61.33	29.18	3	random
25%	49.61	30.28	61.25	29.55	3	mixed
50%	43.16	30.81	53.58	30.07	3	mixed
75%	31.56	31.79	42.57	30.93	3	mixed
100%	0.00	∞	12.97	41.60	3	correct

high-fidelity cover reconstruction without the correct key. Thus, the key-gated mechanism should be interpreted as an authorization boundary for recovery rather than as encryption of the protected image.

Chroma subsampling. The current implementation and controlled experiments use 4:4:4 JPEG, with spatially corresponding 8×8 Y, Cb, and Cr blocks. Subsampling to 4:2:2 or 4:2:0 changes both the available chrominance capacity and the correspondence between luminance and chrominance blocks, so the current allocation and mapping cannot be applied unchanged. In particular, conversion of an already protected image from 4:4:4 to 4:2:0 low-pass filters and downsamples Cb/Cr by two in each spatial dimension. Records embedded separately in neighboring chrominance blocks can therefore be mixed, disrupting the original perturbation–record correspondence. Native subsampled encoding and subsampling conversion after protection require separate design and validation; neither is established by the current controlled experiments or aggregate platform results.

Non-aligned recompression. In a diagnostic experiment, protected images generated at $QF_p = 75$ were cropped by removing three pixels from both the top and left boundaries and then recompressed at different quality factors. The original block boundary at $(8, 8)$ moves to $(5, 5)$, while recompression introduces a competing grid at $(0, 0)$. Both grids can exhibit quantization-integer consistency, making the 64-phase search ambiguous. The diagnostic observations showed stronger interference from the new grid and poorer recovery as the recompression quality approached $QF_p = 75$. Successful synchronization permits recovery on the surviving aligned region; an incorrect phase misaligns Y perturbations and Cb/Cr records, and removed boundary pixels remain outside the recoverable support. Thus, block-phase synchronization does not guarantee recovery after arbitrary non-aligned recompression. Future work could investigate record-aware synchronization and subsampling-aware allocation, which require separate implementation and validation.

N Difference from Robust Reversible Watermarking

Robust reversible watermarking and SRAP-JPEG both involve robustness and reversibility, but the two objectives are different. In many robust reversible watermarking schemes, a robust watermark signal is first embedded to support message extraction under distortion, and reversible data hiding is then used to record the embedding distortion or auxiliary information needed to restore the cover image. The robustness mainly refers to the watermark message: the embedded message can still be detected or extracted after signal processing. The reversible part, however, is usually defined under an intact or nearly intact decoding condition, because the reversible records produced by RDH can be altered, shifted, or removed by recompression, cropping, resizing, or other post-processing operations.

SRAP-JPEG targets a different and stricter requirement. The embedded adversarial perturbations should remain effective before authorization, while the authorized recovery process should also remain usable after practical image-circulation distortions. Thus, the challenge is not only to make the protected JPEG robustly adversarial, but also to preserve or recover the perturbation–record correspondence needed to remove adversarial effects and reconstruct high-fidelity content after the protected image has been modified. This distinction is why SRAP-JPEG combines keyed cross-channel recording, block-phase synchronization, embedding-state identification, and coefficient restoration, rather than relying on clean-bitstream reversible extraction alone.

O Additional qualitative examples.

We provide additional qualitative examples to complement the quantitative analyses in the main paper. Fig. 15 presents representative SRAP-JPEG protected images under JPEG quality factors from 50 to 90 together with their corresponding file-size changes, illustrating the practical storage–quality trade-off. Figs. 11–14 further visualize recovered images after the protected images are subjected to Gaussian noise, JPEG recompression, edge cropping, and resizing, respectively. These examples are consistent with the quantitative robustness results: SRAP-JPEG maintains favorable visual fidelity under a broad range of realistic image circulation distortions, while degradation increases gradually rather than abruptly as the distortion severity becomes stronger.

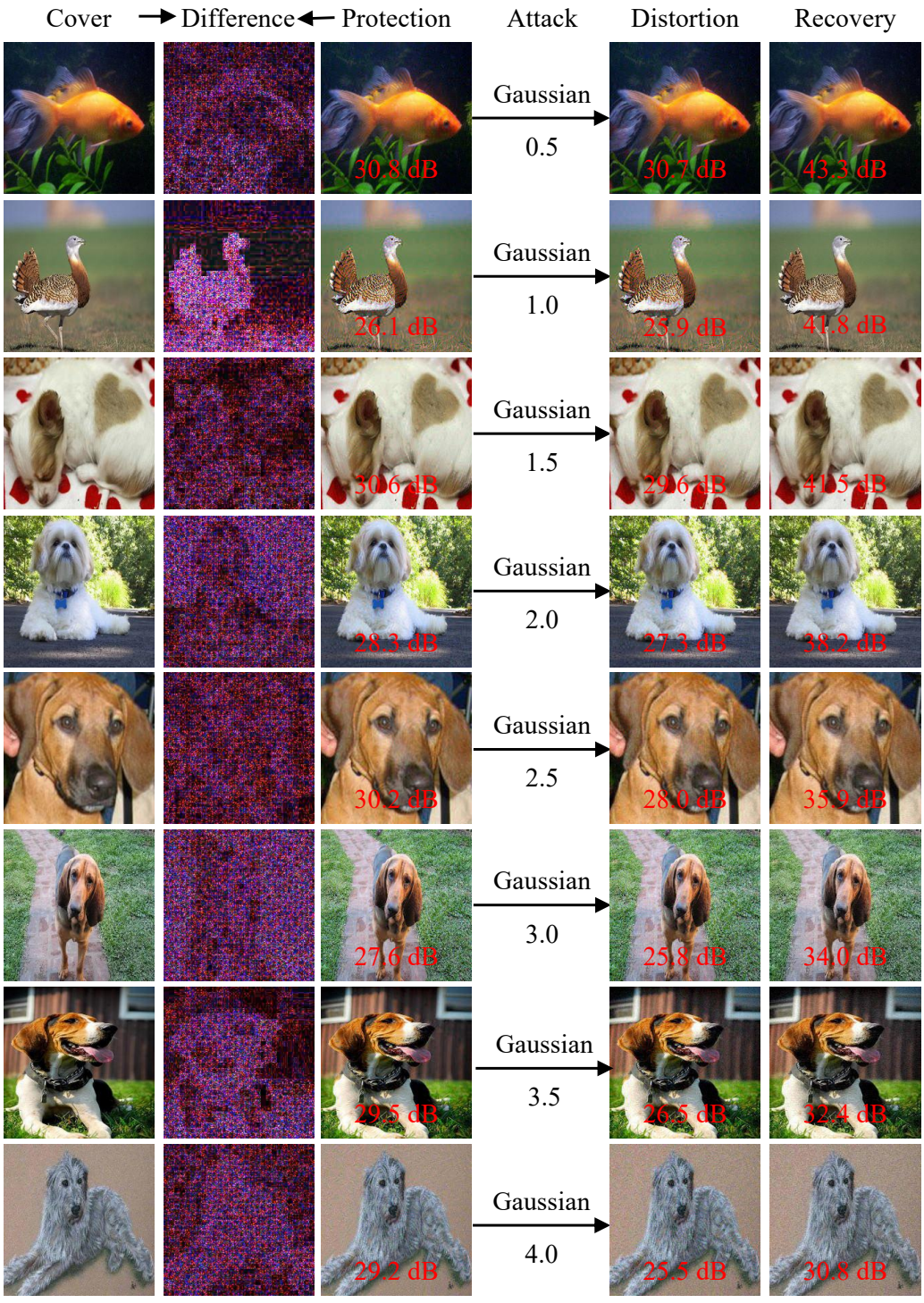

Figure 11: Representative recovery results of SRAP-JPEG under Gaussian noise with severity levels from 0.5 to 4.0. The recovered images remain visually faithful over a wide range of additive noise strengths, with gradually increasing degradation under stronger noise.

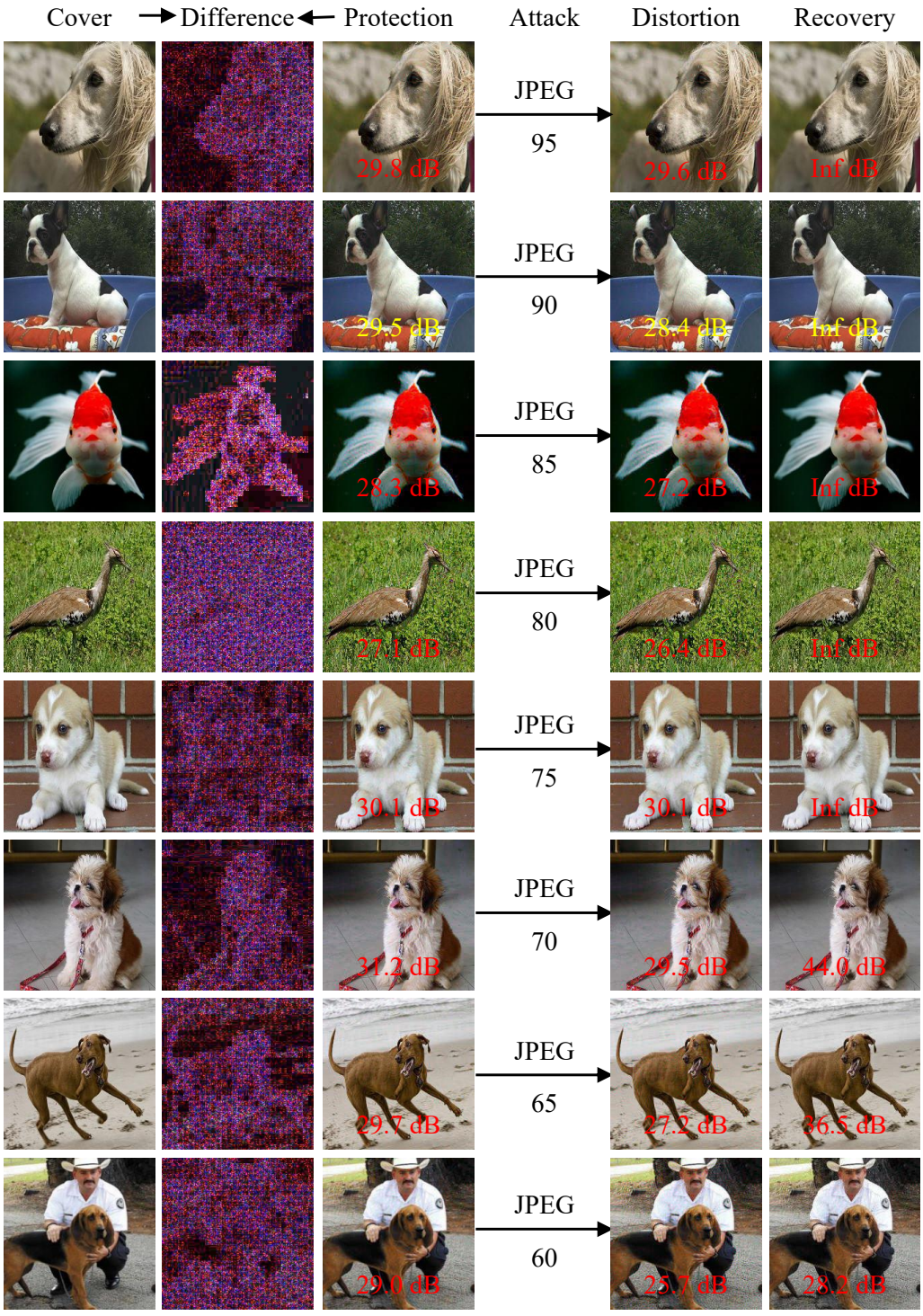

Figure 12: Representative recovery results of SRAP-JPEG under JPEG recompression with quality factors from 95 to 60. The recovered images preserve favorable visual quality under moderate recompression and degrade gracefully as the recompression becomes more aggressive.

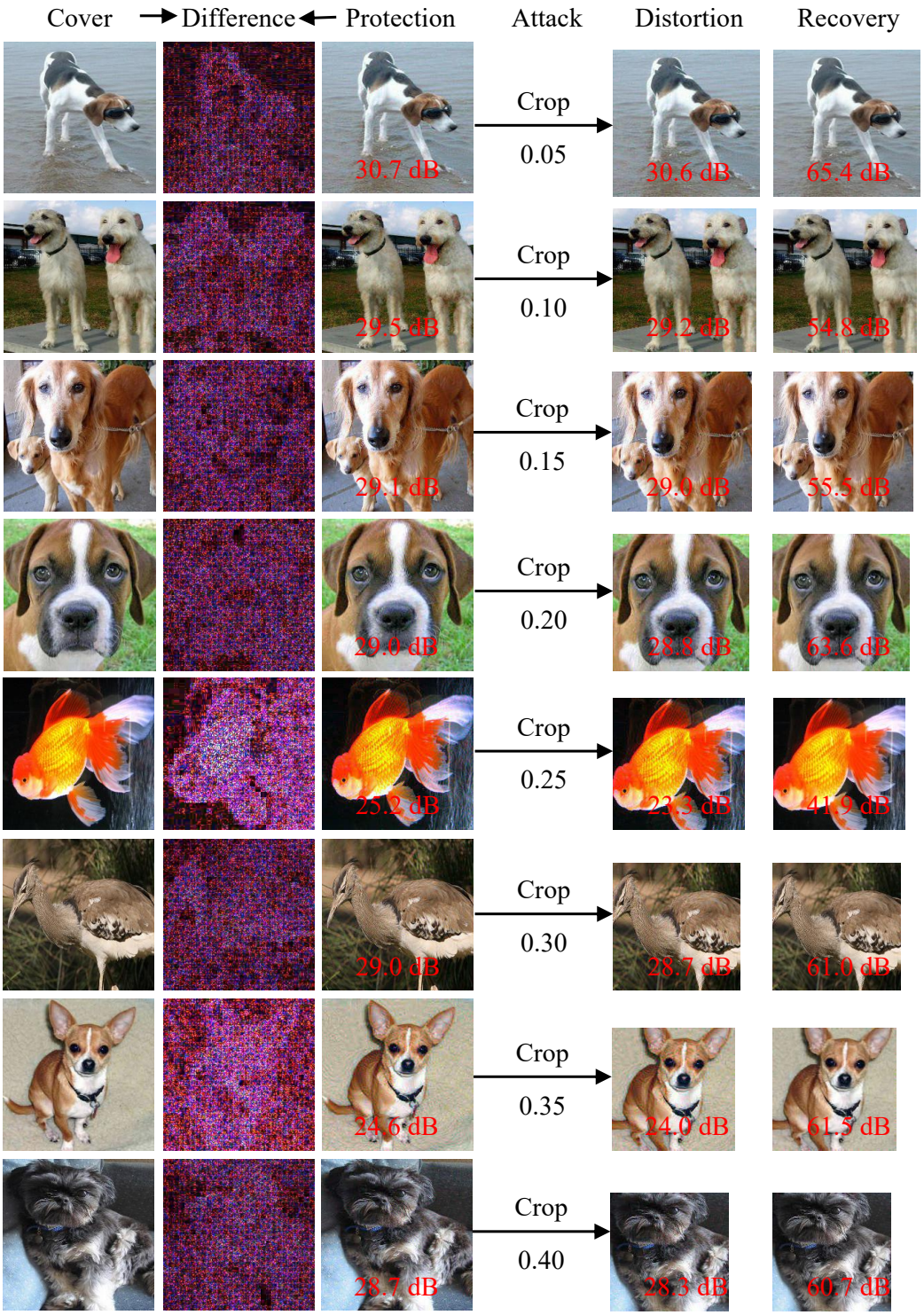

Figure 13: Representative recovery results of SRAP-JPEG under edge cropping with crop ratios from 0.05 to 0.4. Despite the spatial misalignment introduced by cropping, SRAP-JPEG still restores visually consistent content across a broad range of crop severities.

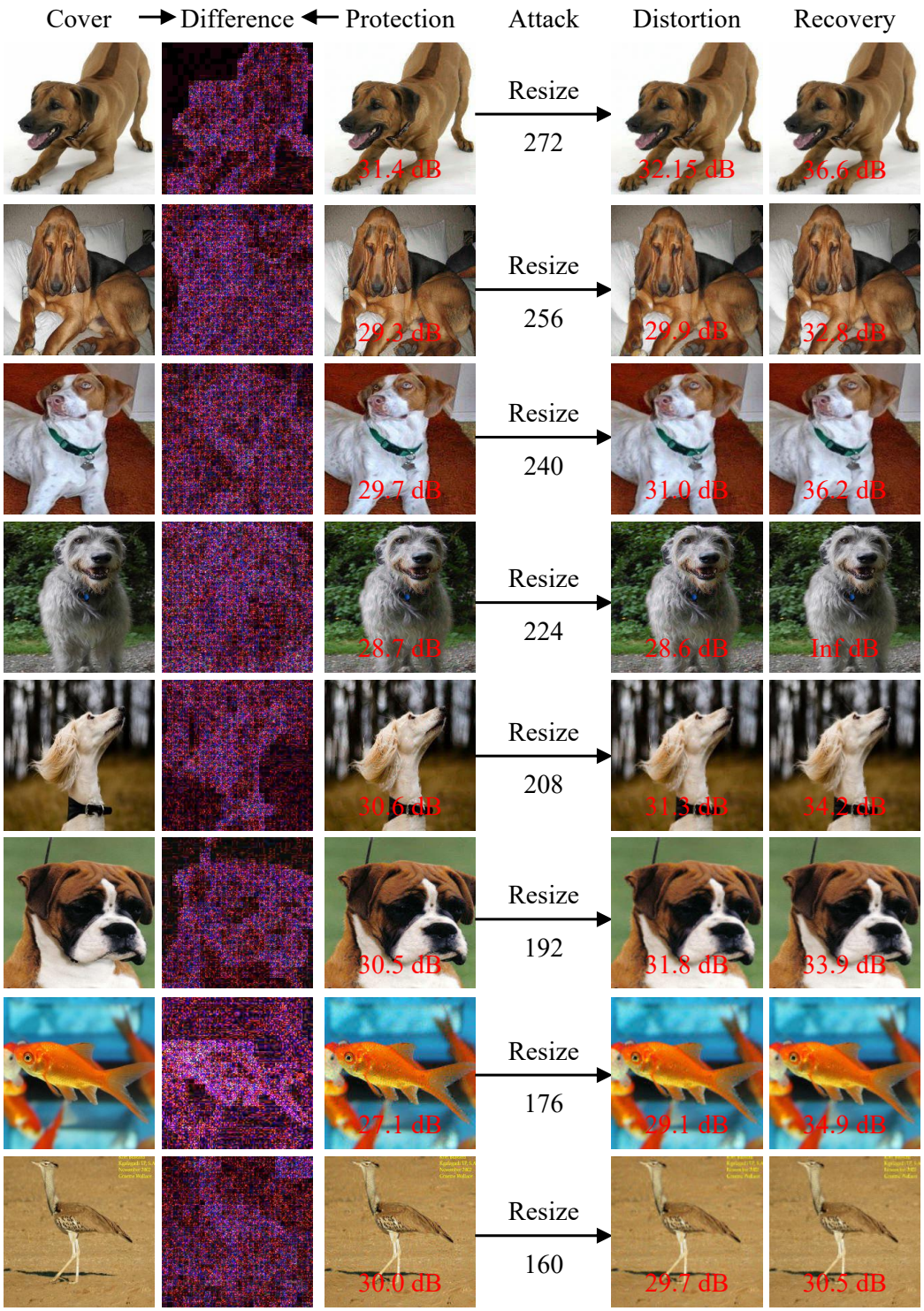

Figure 14: Representative recovery results of SRAP-JPEG under resizing with target sizes from 272 to 160. The recovered images remain visually stable under scale changes, while stronger downsampling leads to only gradual quality loss.

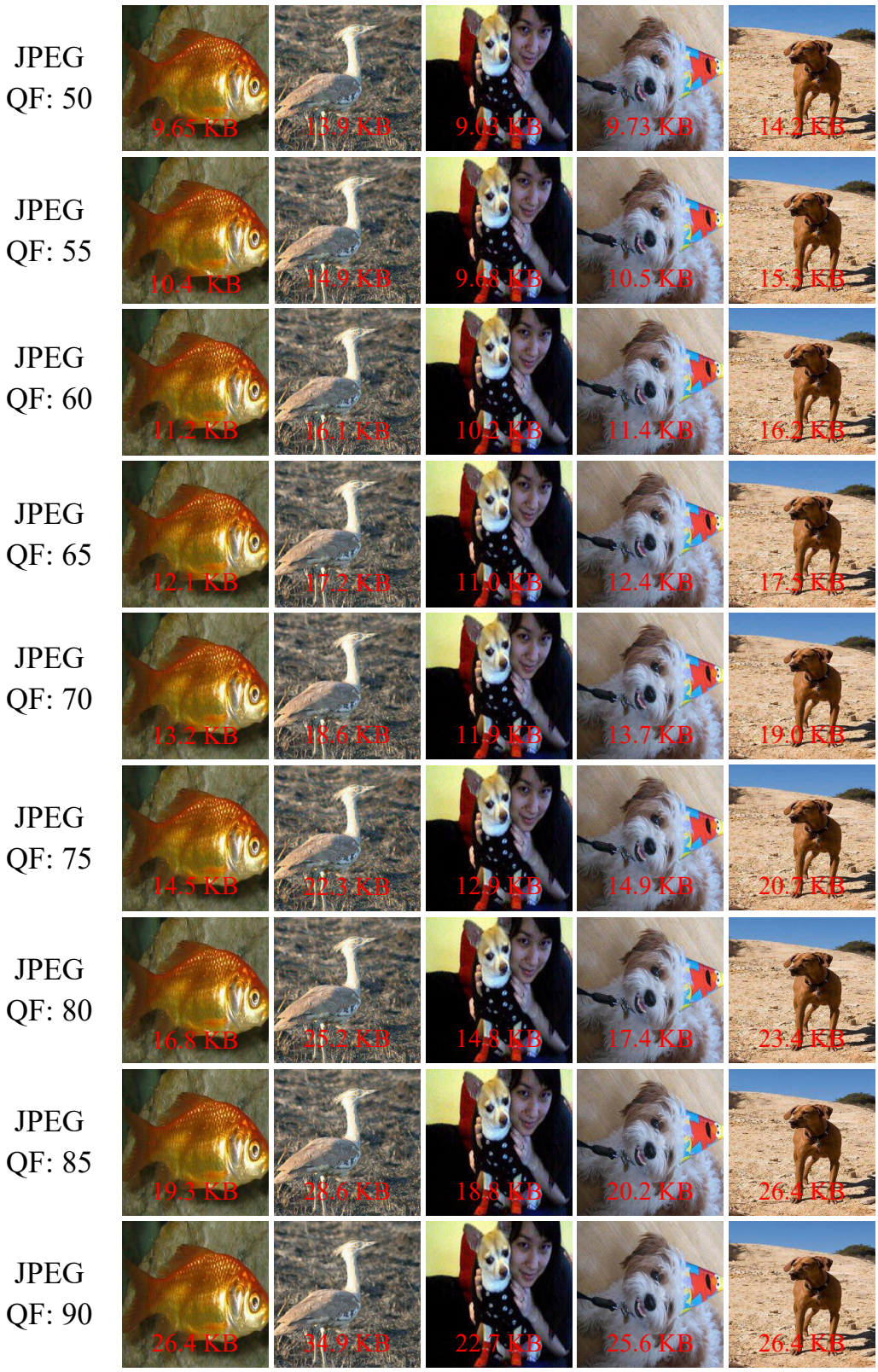

Figure 15: Representative SRAP-JPEG protected images under JPEG quality factors from 50 to 90. The figure also reports the corresponding protected-image sizes, showing how higher compression quality improves visual appearance while increasing storage cost.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of SRAP-JPEG, its JPEG-domain design components, and its empirical claims on adversarial effectiveness, recovery quality, robustness, and file-size growth. These claims are supported by the method description in Section 3, the main results in Section 4, and the detailed analyses in the appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The appendix discusses the scope and limitations of the key-gated recovery mechanism in Appendix L and M. The experiments also delimit the tested setting to public vision datasets, JPEG images resized to 224×224, and the reported image-circulation distortions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theorems requiring proofs. The mathematical expressions define the JPEG-domain protection, gradient propagation, and recovery procedures, and the differentiability argument is explained in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and Appendix D specify the datasets, evaluated model families, preprocessing, JPEG quality factor, SRAP-JPEG settings, baseline configurations, evaluation metrics, and post-processing distortions needed to reproduce the main comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The experiments use public datasets and cited baseline methods, but this draft does not yet provide an anonymized public code release, exact execution commands, or a complete reproduction package.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 describes the benchmark datasets, model families, metrics, and experimental protocol. Appendix D further gives baseline configurations, preprocessing, the default quality factor, threshold τ, batch size, and optimization iterations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports aggregate metrics over the evaluated image sets, but it does not include error bars, confidence intervals, or formal significance tests for the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D reports the GPU model, GPU memory, CUDA version, batch size, average per-image runtime, and estimated runtime for generating protected images on the evaluated datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is framed as a privacy, copyright, and data-protection mechanism, uses standard public vision datasets, and does not involve human-subject experiments, private data collection, or release of high-risk pretrained models. Appendix L clarifies the authorization scope of the key-gated recovery mechanism.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction and conclusion discuss positive uses for privacy, copyright protection, and secure circulation of JPEG images. The appendix also analyzes dataset-protection and poisoning behavior, and Appendix L clarifies that the method supports key-gated recovery rather than cryptographic secrecy, which bounds its deployment claims.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk pretrained model, image generator, scraped dataset, or other asset requiring special release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper cites the datasets, models, metrics, and baseline methods used in the experiments, but this draft does not yet explicitly list the licenses, versions, URLs, or terms of use for all existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, benchmark, pretrained model, or other standalone research asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The experiments do not involve crowdsourcing or human-subject studies.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve human-subject experiments requiring IRB or equivalent review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method or experiments; any writing or formatting assistance does not affect the scientific contribution.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.