IRIS: IMPERCEPTIBLE, ROBUST IMAGE SEALING FOR COOPERATIVE ANTI-EDIT PROTECTION

Anonymous authors
Paper under double-blind review

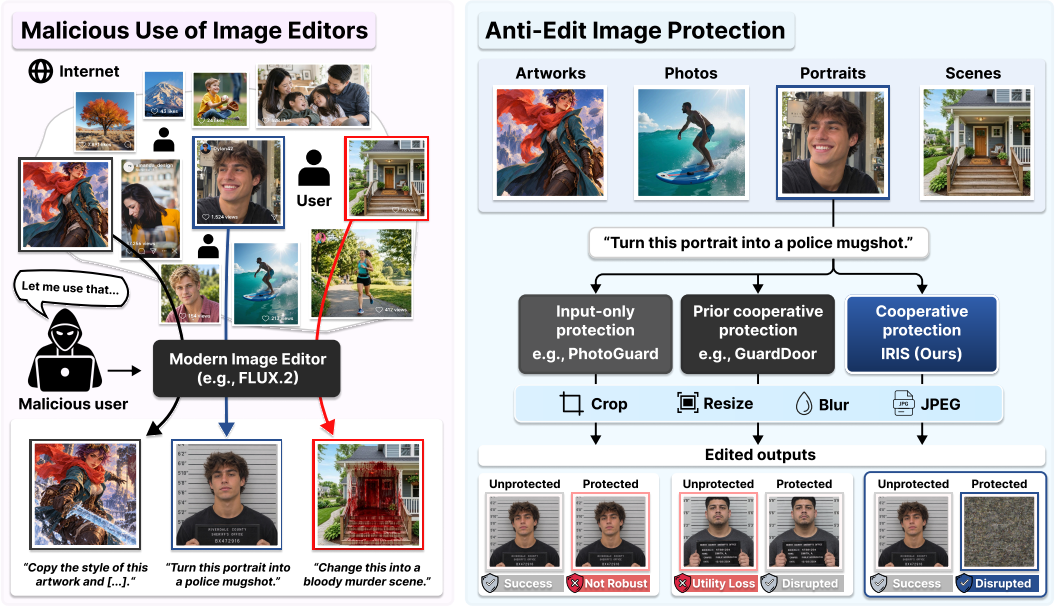


Figure 1: **Left:** Image editors can repurpose public photos and artworks without consent, enabling harm and misinformation. **Right:** Protection must survive image processing. Input-only defenses (Salman et al., 2023; Chen et al., 2025a) can lose robustness; cooperative defenses (Zeng et al., 2025) can impair clean utility. IRIS combines watermarks with learned velocity suppression to disrupt protected edits while largely preserving clean editing.

ABSTRACT

Unified image-editing models such as FLUX.2 unlock substantial creative potential, but also make publicly shared content easy to repurpose without consent. For example, they can fabricate images of a person in scenarios that never occurred or imitate an artist's signature style. Existing input-only defenses add subtle perturbations to disrupt such edits, but this small signal must survive everyday image processing. Simple operations such as cropping, flipping, resizing, or compression can therefore weaken protection, while increasing perturbation strength can compromise visual quality. We introduce IRIS, a *cooperative* defense in which the model provider participates in enforcing the image owner's opt-out. This cooperation allows IRIS to separate robust signal design from protective response learning. It embeds subtle, image-dependent signals using a frozen watermarking model and fine-tunes the editor with a target-image-free objective that selectively *cools down* the velocity field on protected inputs, keeping generation in the noisy regime while preserving clean editing. Creators can apply the protection to new images without per-image optimization or access to the editor, and the same protected image can be used across cooperating models. Experiments with FLUX.2 and QWEN-IMAGE-EDIT demonstrate this selectivity: IRIS combines low perceptual distortion and strong protection under everyday image processing with largely preserved clean-image editing fidelity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Unified generative models can both synthesize new content and convincingly modify existing images through natural-language instructions (Labs, 2025; Wu et al., 2025). This expands their creative potential, but also makes publicly shared images easy to misuse. A person's portrait can be repurposed for a fake dating profile, turned into an advertisement implying an endorsement they never gave, or placed into situations that never occurred. An artist's work can similarly be altered, repurposed, or imitated without their consent or control. Crucially, neither the source nor the requested edit has to appear harmful on its own. A prompt filter can assess the request, but cannot determine whether the image owner consented to it.

Users should therefore be able to express an image-level opt-out from generative editing. Useful protection has to survive routine processing, such as cropping, resizing, and compression. At the same time, it has to also preserve the editing model's utility: model developers need safeguards that restrict unauthorized uses without notably compromising ordinary generation and editing.

Existing approaches address parts of this problem. Input-only defenses add subtle perturbations intended to disrupt an unchanged editing model (Salman et al., 2023; Chen et al., 2025a; Ozden et al., 2026). The same perturbation must survive processing and remain effective afterward, placing both demands on a limited distortion budget. Defenses developed for earlier workflows also struggle to transfer to recent unified editing models (Grebe et al., 2026a). Alternatively, an external watermark detector can reject protected inputs, but this check can be omitted when model weights are released. Parameter-level safeguards such as concept erasure modify the model itself, although balancing robust suppression with preservation of benign capabilities remains challenging (Srivatsan et al., 2025; Huang et al., 2025; Braun et al., 2026). Moreover, concept-level suppression cannot provide a general opt-out mechanism; it can only suppress specific content associated with a given concept.

This motivates a *cooperative* setting in which the provider participates in enforcing the owner's opt-out choice (Zeng et al., 2025). Cooperation separates two responsibilities: the image carries a robust opt-out signal, while the editing model supplies the protective response. The central challenge becomes *selectivity*: learning to suppress edits on protected inputs, including after image processing, while preserving ordinary editing on clean inputs.

We introduce IRIS, a cooperative image-level opt-out for unified editing models. A frozen watermark embedder (Sander et al., 2025) supplies a subtle input signal, while our response-learning objective encourages the editing model's generative velocity to vanish on protected inputs, stalling generation in the noisy regime. On clean inputs, the adapted editor is trained to match the original editor's predictions. Training requires only ordinary source images and self-supervision, without curated editing instructions, ground-truth edited-image pairs, or a predefined target image.

Our evaluation of IRIS on FLUX.2 and QWEN-IMAGE-EDIT jointly examines protection robustness, perceptual distortion, and clean editing fidelity across multiple datasets and inference settings. Creators can protect new images without per-image optimization or access to the editing model. Providers can independently adapt their models to the same signal, allowing one protected image to express an opt-out across multiple cooperating editing models.

Our contributions are:

- We introduce a target-image-free, velocity-suppressing objective that couples selective disruption of protected edits with preservation of clean editing, without requiring a curated editing dataset. We compare alternative response objectives to isolate the roles of velocity suppression and clean-behavior preservation.
- We demonstrate strong protection and robustness with low input distortion: on FLUX.2, protected VLM edit success averages 4.7% across tested common image transformations, while untransformed clean edit success decreases by only two percentage points relative to the original editor. Our evaluation additionally covers adversarial purification and multi-reference generation.
- We show that IRIS naturally supports simultaneous protection against several cooperating models using a single input perturbation, offering a clear advantage over prior input-only and model-specific cooperative baselines.

- We qualitatively demonstrate that localized inference-time amplification can improve resistance to adversarial purification while limiting added perceptual distortion. This gives users further control over the trade-off without changing model parameters.

2 RELATED WORK

Mitigating Generative Model Risks. Early image editors repurposed pretrained diffusion models through inversion or specialized conditioning (Couairon et al., 2023; Wallace et al., 2023; Mokady et al., 2023), while subsequent models learned instruction-driven editing directly such as INSTRUCTPIX2PIX (Brooks et al., 2023; Zhang et al., 2023). More recently, flow-based architectures combine generation and editing within unified systems (Labs, 2025; Wu et al., 2025). Existing safeguards guide generation away from unsafe content (Schramowski et al., 2023) or modify model parameters to suppress specified concepts (Gandikota et al., 2023). Related unlearning work uses noise-like outputs to suppress selected semantic classes (Spartalis et al., 2025). IRIS conditions suppression on an opt-out signal embedded in the source image and uses clean teacher matching to preserve ordinary editing, providing image-level control complementary to content-level restrictions.

Adversarial Perturbations and Image Cloaking. Image cloaking uses small pixel perturbations to protect visual content from downstream exploitation. Early approaches focused on preventing unauthorized learning from personal photos or artworks (Shan et al., 2020; Huang et al., 2021; Shan et al., 2023; 2024; Liang et al., 2023). Later methods target editing misuse by perturbing encoder representations or the generative process (Salman et al., 2023; Chen et al., 2025a; Kim et al., 2025). These approaches optimize protection separately for each image, as does VETO, which extends anti-edit protection to unified models by disrupting reference-image attention (Grebe et al., 2026a). DIFFVAX instead trains an immunizer network once to produce protected images without per-image optimization at deployment (Ozden et al., 2026). These input-only approaches leave the editor unchanged and do not require provider cooperation but simple transformations such as resizing, cropping, or JPEG compression can weaken or remove their protection (Salman et al., 2023).

Image Watermarking. Embedding imperceptible signals into images is a well-established approach to provenance and ownership verification (Cox et al., 1997). Recent methods introduce watermarks through the generative model or sampling process (Fernandez et al., 2023; Wen et al., 2023), or learn image-dependent perturbations that can be added to existing images (Sander et al., 2025; Souček et al., 2025). WAM learns to embed, localize, and recover messages, including from partially watermarked images (Sander et al., 2025). PIXEL SEAL further addresses imperceptibility and robustness through adversarial and message-recovery objectives with staged training (Souček et al., 2025). EDITGUARD combines copyright recovery with tamper localization (Zhang et al., 2024). These methods support signal recovery or identification of manipulation after it occurs but do not themselves help to prevent it. IRIS instead uses pretrained watermark embedders to communicate an opt-out to a cooperating editor. However, robust watermark detection alone does not establish robust downstream editing protection.

Cooperative Protection and Conditional Responses. A small number of approaches assume that the model itself can be modified in support of input-level protection. GUARDDOOR (Zeng et al., 2025) follows this cooperative setting: it uses model-specific VAE reconstruction patterns as triggers and fine-tunes the image encoder to map protected inputs to a prescribed latent target while preserving clean behavior. The closely related approach of Chen et al. (2025b) uses pretrained watermark signals to condition fixed-target responses in INSTRUCTPIX2PIX, with paired editing supervision and image-reconstruction losses. IRIS instead couples direct-zero velocity regression with clean teacher matching in modern flow-based unified editing models, using empty training instructions and no ground-truth edit pairs or prescribed output image. VISILOCK (Le & Fu, 2026) considers a complementary setting, enabling editing only when a visual key is present. This implements access control for the editing model rather than an image-specific opt-out. Such a mechanism is of limited practical use for open-source models, since a model that requires a visual key for ordinary editing would substantially restrict its general usability; in closed-source settings, meanwhile, access can be controlled more directly through the model or service itself.

Our evaluation shows a better overall protection–utility trade-off than GUARDDOOR under common image transformations. Objective controls within a shared training protocol and signal–response ablations examine the contributing design choices. We further demonstrate that the same protected

image supports opt-out behavior across independently adapted editors and disrupts multi-reference editing when only one reference carries the signal.

3 METHOD

IRIS combines an imperceptible input signal with a protective response learned by a cooperating editing model, which we often refer to as the *editor* for brevity (Figure 2). A frozen embedder supplies the robust signal, while our learning of the selective response teaches the editor to suppress protected edits and preserve ordinary clean editing.

Threat Model and Defense Setting. Before publishing an image x, its owner applies a protection signal to obtain x_P. A malicious user attempts to edit x_P with a publicly available editor f and arbitrary instruction p, potentially after transforming the image to weaken protection. We assume that the provider *cooperates* by fine-tuning f to suppress protected edits while preserving clean editing before releasing the model f. We evaluate protection under continued use of the adapted editor. Multiple providers can independently adapt their editors to the same signal. We assume neither control over downstream inference nor secrecy of the embedder, message, or defense mechanism.

Open-weight editors can run locally or in third-party applications, beyond provider-enforced safeguards. Embedding the response in released weights allows the adapted editor to enforce the opt-out outside a provider-controlled interface. Our objective preserves ordinary editing while suppressing protected edits, making selectivity central to practical adoption (Appendix A.1).

We first describe how the signal is embedded, then how the editor learns to respond to it. Algorithm 1 in Appendix B.9 summarizes training.

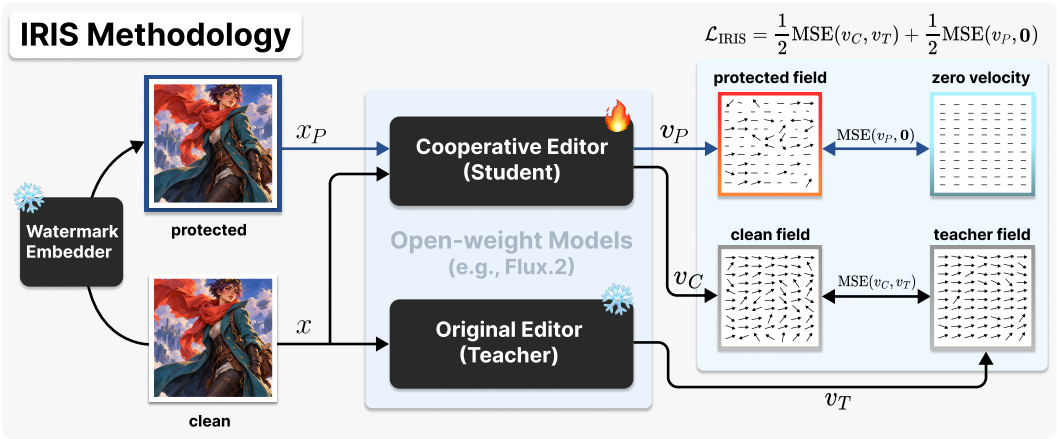


Figure 2: IRIS response training. A frozen pretrained watermark embedder produces protected counterparts of clean inputs. At matched noisy latents, timesteps, and empty instructions, a LoRA-adapted student predicts clean and protected velocities v_C and v_P, while the frozen original editor supplies detached teacher predictions v_T. The objective $\mathcal{L} = \frac{1}{2}\,\text{MSE}(v_C, v_T) + \frac{1}{2}\,\text{MSE}(v_P, \mathbf{0})$ preserves clean-input dynamics while cooling down protected-input velocities to suppress generation.

3.1 PROTECTION-SIGNAL EMBEDDING

For a clean source image $x \in [0,1]^{3\times H\times W}$ and a fixed binary message m, let $W_\phi(x, m)$ denote a pretrained watermark embedder. We construct

$$x_P = Q\big(W_\phi(x, m)\big), \qquad \delta_x = x_P - x, \tag{1}$$

where Q clips to the image range and rounds to 8-bit RGB. The message m is fixed, but the residual δ_x remains image-dependent. With ϕ frozen, protection requires one embedding pass, without further watermark training, per-image optimization, or editor access. A fixed message does not imply message-specific recognition; Appendix C.4 examines this distinction.

The signal should introduce little change while remaining informative after image transformations. Robust watermarking directly addresses these requirements. We use WAM, which learns spatial watermark recovery under masking and image distortions and applies perceptual attenuation to reduce visible artifacts (Sander et al., 2025). We also examine PIXEL SEAL as an alternative, motivated by its adversarial image-quality training and reported watermark robustness (Souček et al., 2025).

To vary signal strength, we also evaluate $Q(x + \alpha \cdot \delta_x)$ for $\alpha > 0$, where $\alpha = 1$ gives the native protected image. Our main WAM configuration uses $\alpha = 1$. Appendix C.2 examines this trade-off. Below, x_P denotes the protected input at the selected strength. The same response objective can accommodate different embedders, but watermark recovery alone does not establish editing protection. The editor must learn a selective response to the surviving signal.

3.2 SELECTIVE RESPONSE LEARNING

We fine-tune the denoising transformer through teacher-student distillation (Hinton et al., 2015; Gandikota et al., 2023), preserving clean-input velocities while learning to suppress protected generations. The original model is the frozen teacher and the adapted one the student, with W_ϕ fixed throughout. GUARDDOOR instead adapts the image encoder; Appendix A.4 compares these choices.

Our central design is a selective *cooldown* of the velocity field. Rather than teaching the editor to generate a prescribed target image, we train its velocity predictions toward zero whenever the protection signal is present. In a flow-based editor, these predictions determine how the noisy latent evolves toward the generated image. Suppressing them therefore stalls this evolution, leaving the output insufficiently denoised. An exactly zero effective velocity field would halt the flow; our objective encourages this behavior without requiring exact suppression at every sampling step. On clean inputs, distillation anchors the generative dynamics to those of the original editor.

Intuitively, learning to suppress generation may be simpler than learning to synthesize a prescribed replacement image, potentially easing selective adaptation. Our ablations support this design choice: the shared objective achieves a better protection–utility balance than the tested alternatives on both editors (Appendix A.2). Appendix A.3 further examines the resulting velocity suppression.

Let z_x denote the original editor's latent endpoint for source x under the empty instruction. With $t \sim \mathcal{U}[0, 1]$ and $\xi \sim \mathcal{N}(0, I)$, we form

$$z_t = (1 - t)z_x + t\xi. \tag{2}$$

Let g_{θ^*} denote the frozen teacher and g_θ the adapted student. At the latent state (z_t, t), with empty instruction $p = \varnothing$, we get the teacher, clean-input student, and protected-input student velocities:

$$v_T = g_{\theta^*}(z_t, t; p, x), \quad v_C = g_\theta(z_t, t; p, x), \quad v_P = g_\theta(z_t, t; p, x_P). \tag{3}$$

The teacher prediction v_T is detached from the computation graph, while gradients pass through both student predictions. Training requires only a pool of images, without curated editing instructions, ground-truth edited-image pairs, or supervision from a watermark decoder.

We combine clean-behavior distillation with protected-velocity suppression:

$$\mathcal{L}_{\text{IRIS}} = \underbrace{\tfrac{1}{2}\,\text{MSE}(v_C, v_T)}_{\text{clean preservation}} + \underbrace{\tfrac{1}{2}\,\text{MSE}(v_P, \mathbf{0})}_{\text{velocity suppression}} \tag{4}$$

Averaged over source images, times, and noise, the equally weighted terms teach the same model to preserve ordinary editing on clean inputs and stall generation when the protection signal is present.

At deployment, owners embed protection once, and cooperating editors respond through their ordinary inference pipeline without an external detector or rejection check. We fine-tune with LoRA Hu et al. (2022), whose updates can be merged into the released weights, so protection does not reside in a separate adapter that users can simply disable.

4 EXPERIMENTS

We primarily evaluate IRIS on FLUX.2 (Labs, 2025) and QWEN-IMAGE-EDIT (Wu et al., 2025), independently adapting both editors to the same frozen watermark signal. A smaller-scale study

on FLUX.1 KONTEXT (Labs et al., 2025) further demonstrates applicability to an additional flow-matching editing backbone (Appendix E.4).

Baselines. We adapt the cooperative baseline GUARDDOOR (Zeng et al., 2025) to both primary editing models, retaining its VAE reconstruction signal and encoder fine-tuning. On FLUX.2, we additionally compare the input-only defenses PHOTOGUARD, EDITSHIELD, and VETO. Further details on the baselines and on objective and component ablations examining alternative targets and signal-response combinations are provided in Appendix B.2.

Evaluation. We use all 300 source images and native editing instructions from VETOBENCH (Grebe et al., 2026a). DINO cosine similarity (Caron et al., 2021) compares outputs with unprotected (clean) original edits under matched instructions, transformations, and seeds. High clean-edit similarity indicates preservation; low protected-edit similarity indicates disruption. An output can differ substantially from the original editor's result yet still satisfy the editing request, motivating a complementary assessment of editing success. Therefore, we additionally report VLM-judged editing success using QWEN3.5-27B (Team, 2026). Given the instruction, actual editor input, and edit output, the judge evaluates instruction compliance, a coherent and useful result, and appropriate source correspondence. An edit counts as successful only when all three criteria are satisfied. Moreover, on 400 General-category cases, VLM judgments agree with three independent human reviewers in 87.0%–90.3% of comparisons and preserve their method ranking (Appendix B.6). Input LPIPS and PSNR measure protection-induced distortion. To validate that IRIS transfers to real-photo distributions, we additionally test on 100 MagicBrush samples (Zhang et al., 2023) (Appendix E.3). Appendix B.10 also evaluates changes to guidance and sampling duration without retraining.

Robustness and Multi-Reference Editing. We evaluate untransformed inputs, horizontal flipping, weak and strong JPEG compression, downscaling, cropping, and blurring (Appendix B.4). Our JPEG settings (Q70/Q40) use more aggressive compression than the Q80 reported by GUARDDOOR and EditShield (Zeng et al., 2025; Chen et al., 2025a). We also test DiffPure at two noise levels ($t = 50, 150$), with matched purified clean references. Appendix D details the purification evaluation and differences from prior protocols.

For multi-reference editing, we evaluate FLUX.2 on 50 MultiBanana (Oshima et al., 2026) tasks: 25 with two and 25 with three references. Exactly one randomly selected reference is protected per task, fixed across methods. We report matched-reference DINO similarity and qualitative examples.

Training. We train on OpenImages sources (Kuznetsova et al., 2020) using empty instructions and original-editor supervision, with a frozen watermark embedder and no image augmentation or curated edit pairs. Evaluation uses the benchmark editing instructions, testing whether the response learned with empty instructions remains selective under instruction-guided editing. Objectives and training durations are selected on separate development sources, considering clean preservation and protected disruption jointly. Training and inference settings are provided in Appendix B.8.

5 RESULTS

We assess protection success together with clean utility and imperceptibility. In the main comparisons, IRIS uses WAM (Sander et al., 2025); alternative signals are examined in Appendix A.4.

5.1 SELECTIVE PROTECTION AND PRESERVED UTILITY

On FLUX.2, IRIS retains 88.3% clean VLM edit success, only two percentage points below Original's 90.3%, while protected success falls to 0.3% (Table 1). GUARDDOOR achieves 44.3% clean and 18.0% protected success. Thus, IRIS combines stronger protection with substantially better preservation of clean-image editing. DINO supports this separation: clean outputs retain a similarity of 93.9 to their Original references, whereas protected outputs fall to 2.0.

This separation must also survive routine image processing. Across our tested weak and strong ordinary transformations, IRIS averages 4.7% protected edit success, compared with 34.1% for GUARDDOOR and 63.3–86.9% for the input-only defenses. Figure 3 illustrates disrupted protected outputs alongside a clean edit that follows the instruction. Strong JPEG compression (Q40) remains the most challenging condition, with 33.0% protected success; the other ordinary conditions yield 0.3–3.7%. Appendices E.1 and E.7 provide the complete breakdown and further examples.

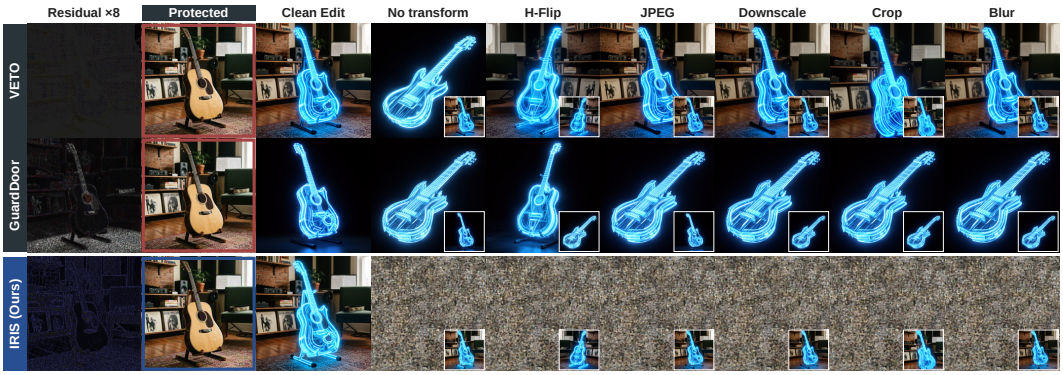


Figure 3: Protection comparison on FLUX.2 - Instruction: "Convert this instrument into a futuristic hard-light construct made of glowing neon blue energy lines." "'latex After the Clean Edit column, large panels show protected-input outputs; insets show the same editor's clean-input output with matched transformation, instruction, and seed. IRIS retains the clean edit while protected outputs remain visibly disrupted under the shown transformations.

Table 1: FLUX.2 on 300 VetoBench sources per condition. DINO is matched-output cosine similarity ($\times 100$); VLM is editing success (%). Clean columns use untransformed inputs. Protected transform means average nine ordinary non-identity conditions, excluding DiffPure. IRIS uses WAM; $0.5\times$ is a separately adapted editor trained and evaluated at half signal strength. Input-only clean DINO is a self-comparison. Bold indicates the best value within each defense group.

Method	Input LPIPS ↓	Clean ↑		Protected Edit ↓ No transform		Protected Edit ↓ Mean transform	
		DINO	VLM	DINO	VLM	DINO	VLM
Original	0.000	100.0	90.3	100.0	90.3	100.0	88.3
Input-only defenses							
PHOTOGUARD	0.224	**100.0**	**90.3**	69.4	39.0	94.1	85.1
EDITSHIELD	0.170	**100.0**	**90.3**	72.2	43.3	**82.7**	**63.3**
VETO	**0.054**	**100.0**	**90.3**	**57.7**	**17.7**	96.8	86.9
Cooperative defenses							
GUARDDOOR	**0.009**	66.5	44.3	57.7	18.0	63.7	34.1
IRIS (Ours)	0.038	**93.9**	**88.3**	2.0	0.3	6.8	4.7
IRIS (Ours), $0.5\times$	0.010	83.9	83.3	**1.6**	**0.0**	**2.6**	**0.9**

Native IRIS introduces input LPIPS of 0.038, below the evaluated input-only defenses (0.054–0.224), but above GUARDDOOR (0.009) on FLUX.2. Halving the signal strength during training and inference reduces LPIPS to 0.010, approaching GUARDDOOR while retaining substantially stronger protection and clean editing utility. Across ordinary transformations, protected edit success falls to 0.9%, compared with 4.7% at native strength and 34.1% for GUARDDOOR. This improvement comes with reduced clean edit success, from 88.3% to 83.3%, which remains well above GUARDDOOR's 44.3%. Appendix C.1 examines this trade-off in detail.

Provider-side utility extends beyond reference-based editing. On a 1,000-prompt MJHQ subset, we evaluate reference-free text-to-image generation, which gives a comparable FID of 63.12 to the original model's 62.89. This indicates little change in this distributional quality measure, although it does not imply identical generations. Appendix C.3 examines preservation more fully.

The clean/protected separation also extends beyond the synthetic benchmark: on 100 real MagicBrush photos, our IRIS–FLUX.2 checkpoint retains 85.0% clean VLM success versus Original's 86.0%, with 0.0% of edit success for protected images (Appendix E.3).

IRIS objective ablation. Holding WAM and the training protocol fixed, we ablate our selective response objective against removing the preservation term, maximizing prediction separation (Chen

et al., 2025a), and regressing velocity toward a fixed gray image (Salman et al., 2023; Zeng et al., 2025). Across the four transformations in Table 2, IRIS achieves model-averaged clean/protected DINO of 89.2/1.3, compared with 85.6/84.5 for the fixed-gray control. Removing preservation or using max-separation breaks clean utility. Component swaps further highlight the importance of response learning: the same WAM signal with GUARDDOOR's encoder-response recipe does not reproduce IRIS's selectivity. Appendix A.2 provides further details. Overall, several alternatives suppress protected outputs only by degrading clean editing, showing why disruption alone is insufficient.

Protection vs. imperceptibility trade-off. Purification highlights why robustness must be assessed together with clean utility (Appendix D, Tables 14 and 15). At strong DiffPure ($t = 150$), the input-only defenses largely lose their protection, while GUARDDOOR's low protected edit success is accompanied by similarly low clean success: 35.0% versus 34.3% on FLUX.2. GUARDDOOR thus disrupts editing without preserving a useful clean/protected distinction. Native-strength IRIS also loses most protection at this setting, but retains substantially higher clean utility (82.3%). At the weaker setting ($t = 50$), IRIS retains partial separation, with clean/protected success of 84.3/80.0% on FLUX.2 and 86.3/75.7% on QWEN-IMAGE-EDIT. These results characterize IRIS at its native operating point; its protection strength α can also be increased at inference time without modifying the editor or its clean behavior.

Table 2: IRIS objective ablation on VetoBench. DINO ($\times 100$) averages horizontal flip, JPEG Q70, crop retaining 75% area, and blur with $\sigma = 1.25$. More details in Appendix A.2.

Ablation	**FLUX.2**		**Qwen-Image-Edit**	
Objective	Clean ↑	Prot. ↓	Clean ↑	Prot. ↓
DIRECT ZERO (IRIS)	**93.6**	**3.0**	84.7	**-0.4**
W/O PRESERVATION	1.7	1.7	-0.2	-0.2
MAX. SEPARATION	28.9	6.2	4.2	4.9
FIXED TARGET (GRAY)	78.5	77.5	**92.7**	91.4

This enables a direct inference-time trade-off between imperceptibility and protection strength (Appendix C.2). On 100 examples, doubling IRIS's inference-time signal reduces protected DINO after DiffPure ($t = 50$) from 90.1 to 8.3 on FLUX.2 and from 73.9 to 0.6 on QWEN-IMAGE-EDIT, without retraining. LPIPS rises from 0.042 to 0.113, making the cost in input distortion explicit. Because only protected inputs change, clean-input behavior remains unchanged. Amplification could also be applied to other signals, but cannot recover clean utility already lost through editor adaptation; it is therefore complementary to, rather than a substitute for, selective response learning.

For improved imperceptibility, strength can further be allocated according to image content. We qualitatively test a smoothed texture mask that concentrates amplification in detailed regions. The illustrated guitar remains disrupted at both $t = 50$ and $t = 150$, at lower LPIPS than the displayed stronger global alternative. This motivates content-dependent strength control, while the quantitative global-scaling study already establishes a controllable protection–distortion trade-off.

5.2 ONE PROTECTED IMAGE ACROSS COOPERATING EDITORS

Separating the signal from the response lets an owner protect an image once for different cooperating editors, each learning its response independently. GUARDDOOR instead derives its signal from each editor's own VAE reconstruction, tying the protected image to a particular model (App. B.2). The evaluated input-only baselines additionally require per-image optimization against the target editor.

We test this shared-signal setting with identical perturbations across FLUX.2 and QWEN-IMAGE-EDIT for all 300 sources and ten conditions. Alongside the FLUX.2 results above, QWEN-IMAGE-EDIT retains 86.0% clean VLM success versus Original's 84.0%, with 0.0% protected success observed in every untransformed and ordinary-transform condition. No model-specific image variants or combined perturbations are needed for these models; the required adaptation takes place on the provider side. Appendix E.2 gives the paired evaluation.

5.3 PROTECTION WITH MULTIPLE REFERENCES

Multi-reference editing tests whether protection remains active alongside clean references that contribute to the requested edit. We use 25 tasks with two and 25 tasks with three reference images from MultiBanana in order to demonstrate edit protection when only one of the references carries the opt-out signal. The full MultiBanana results in Appendix E.6 shows that IRIS retains a pronounced

clean/protected separation. Protected DINO falls to 2.2 and 4.5 for two and three references, while all-clean requests retain an overall similarity of 87.6.

GUARDDOOR yields protected similarities of 60.1 and 61.2, alongside lower overall clean similarity of 64.2. Thus, supplying additional clean references does not restore ordinary output similarity in these IRIS evaluations. Figure 4 illustrates the response for hairstyle transfer and subject insertion.

6 LIMITATIONS AND FUTURE WORK

IRIS relies on provider cooperation and continued use of the adapted editor. Embedding the protective response in model weights removes reliance on an external detection check. Publicly available weights also permit attempts to weaken protection through further fine-tuning, while inversion-based attacks may offer alternative routes to circumvention. Future work should evaluate these settings and investigate how to retain the selective response under model modifications and adaptive inference procedures without compromising ordinary editing.



Figure 4: Two-reference FLUX.2 edits with one protected reference. The highlighted reference shows the IRIS-protected input. Insets show matched all-clean edits. Top: Change hair of the person in *R1* to match the hair in *R2*. Bottom: Add the man from *R2* on the armchair in *R1*.

Although protection survives common image transformations, diffusion-based purification can restore editing capability at native signal strength. Signal amplification can improve resistance, but introduces additional perceptual distortion. This motivates further study of adaptive signal-removal attacks and the image-quality costs of restoring editing capability. Our localized signal amplification provides an initial exploration of this trade-off without retraining the editor. Broader evaluation across images, editing models, and attack settings is needed to establish when image-dependent amplification improves robustness at an acceptable perceptual cost.

7 CONCLUSION

We introduced IRIS, a cooperative defense for unified image-editing models that separates robust input signaling from selective response learning. A frozen watermark embedder communicates an image-level opt-out, while the adapted editor learns to suppress generative velocity on protected inputs and preserve ordinary editing through original-editor supervision. This adaptation requires only source images, without curated instructions or ground-truth edit pairs.

Experiments with FLUX.2 and QWEN-IMAGE-EDIT demonstrate a strong protection–utility balance under the tested common image transformations. Protection requires neither per-image optimization nor message decoding during editing, and the same protected image supports opt-out behavior across independently adapted editors. These results highlight selectivity as a central requirement: effective protection should disrupt unwanted edits while retaining the model's usefulness on clean generations.

We envision protection fine-tuning as a complementary post-training safeguard that incorporates image-owner preferences alongside content-level safety policies. An otherwise benign editing request may still conflict with an owner's wishes, making image-level opt-outs valuable beyond restrictions on predefined harmful concepts. This orthogonal policy dimension motivates layered defenses that combine IRIS with content-level safeguards and provenance mechanisms. Although protection depends on continued use of cooperating editors, IRIS provides a practical foundation for extending such preferences beyond provider-controlled interfaces and reducing future risks of unauthorized manipulation, impersonation, and reputational harm.

AI USE STATEMENT

The central research ideas and methodological decisions originated with the authors. We used generative AI tools to assist with coding and method implementation within frameworks developed by the authors, translating the authors' specifications into executable experiments, preparing and launching experimental runs, formatting tables, supporting analysis, discovery and research, drafting sections, polishing writing, and managing cluster jobs. The generative models and vision-language-model evaluator used as components of our experiments are described separately in Section 4 and Appendices B.1 and B.7. The authors reviewed and validated the AI-assisted work and take responsibility for all research decisions, implementation, experimental validity, interpretation of results, and the final manuscript, including all AI-assisted content.

ETHICS STATEMENT

IRIS aims to help image owners express an opt-out from generative editing and to help cooperating model providers respect that preference. Protection depends on use of the adapted editor and may be weakened by signal removal or further model modification. Our results do not establish that harmful editing or identity misuse is prevented in all settings. The learned response does not authenticate image ownership or the watermark payload, and absence of a signal does not establish consent. Third parties could therefore apply the signal to restrict legitimate editing, while unintended activation could impair benign use. Deployment should communicate these limitations and assess unintended suppression alongside protection effectiveness. Our experiments use existing image datasets, including synthetic VetoBench images and real OpenImages and MagicBrush photographs. Human validation involves three independent participants assessing editing outcomes on 100 VetoBench sources. This limited audit and our automated metrics do not establish effectiveness across demographic groups or all forms of misuse. We report clean-edit utility, input distortion, and circumvention limitations to make these trade-offs explicit.

REPRODUCIBILITY STATEMENT

To facilitate further research and reproducibility, we will publicly release our implementation upon acceptance. Section 3 specifies protection-signal construction and the response-learning objective. Appendix B documents the pretrained models, baseline adaptations, image transformations, and evaluation metrics. Appendix B.8 provides the training data, optimizer and LoRA settings, training seed, compute infrastructure details, and development-based selection criteria. Section 4 and Appendix B describe the evaluation protocol and matched inference settings. Appendix B.7 gives the complete VLM evaluation rubric, and Appendix D details diffusion purification and its matched clean controls.

REFERENCES

Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. WAVES: Benchmarking the robustness of image watermarks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1456–1492. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/an24a.html.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Tobias Braun, Jonas Henry Grebe, Patrick Mohr Gordillo, Marcus Rohrbach, and Anna Rohrbach. Erased but not forgotten: How backdoors compromise concept erasure. In *Forty-third International Conference on Machine Learning*, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 126–142, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-73036-8.

Yu-Feng Chen, Tzuhsuan Huang, Pin-Yen Chiu, and Jun-Cheng Chen. Invisible backdoor triggers in image editing model via deep watermarking. In *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, pp. 1–6. IEEE, August 2025b. doi: 10.1109/avss65446.2025.11149824. URL http://dx.doi.org/10.1109/AVSS65446.2025.11149824.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *International Conference on Learning Representations*, 2023.

I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12):1673–1687, December 1997. ISSN 1941-0042. doi: 10.1109/83.650120. URL http://dx.doi.org/10.1109/83.650120.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22466–22477, October 2023.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, October 2023.

Jonas Grebe, Hossein Shakibania, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. Veto: Towards protecting images from frontier ai editing, 2026a. URL https://arxiv.org/abs/2607.27292.

Jonas Henry Grebe, Tobias Braun, Anna Rohrbach, and Marcus Rohrbach. GEM: Geometric erasure by contrastive velocity matching in rectified flows. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=NBMCwxTRSA.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 360–376, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73661-2.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *ICLR*, 2021.

Jinsu Kim, Yunhun Nam, Minseon Kim, Sangpil Kim, and Jongheon Jeong. Blurguard: A simple approach for robustifying image protection against ai-powered editing. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 28664–28706. Curran Associates, Inc., 2025. doi: 10.52202/085713-0964. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/29806b2e4dcb35e42abed21d43e2b7d8-Paper-Conference.pdf.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. URL https://arxiv.org/abs/2412.03603.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, Jul 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL https://doi.org/10.1007/s11263-020-01316-z.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Van Thanh Le and Yun Fu. Visilock: Authorizing instruction-based image editing with dual score distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15710–15718, June 2026.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047, June 2023.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 448–460, June 2026.

Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep Chinchali, and James Rehg. Diffvax: Optimization-free image immunization against diffusion-based editing. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 83–99, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/0054c8a2c903d46496ca1f98d1322988-Paper-Conference.pdf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Tom Sander, Pierre Fernandez, Alain Oliviero Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 79569–79599, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c5ee1911dffe4886367eee7fca314235-Paper-Conference.pdf.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531, June 2023.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1589–1604. USENIX Association, August 2020. ISBN 978-1-939133-17-5. URL https://www.usenix.org/conference/usenixsecurity20/presentation/shan.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. Glaze: Protecting artists from style mimicry by Text-to-Image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 807–825. IEEE, May 2024. doi: 10.1109/sp54263.2024.00207. URL http://dx.doi.org/10.1109/sp54263.2024.00207.

Tomáš Souček, Pierre Fernandez, Hady Elsahar, Sylvestre-Alvise Rebuffi, Valeriu Lacatusu, Tuan Tran, Tom Sander, and Alexandre Mourachko. Pixel seal: Adversarial-only training for invisible image and video watermarking, 2025. URL https://arxiv.org/abs/2512.16874.

Christoforos N. Spartalis, Theodoros Semertzidis, Petros Daras, and Efstratios Gavves. Unleashing uncertainty: Efficient machine unlearning for generative ai, 2025. URL https://arxiv.org/abs/2508.20773.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M. Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23765–23774, June 2025.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22532–22541, June 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Hanyi Wang, Han Fang, Yupeng Qiu, Shilin Wang, and Ee-Chien Chang. Resguard: Enhancing robustness against known original attacks in deep watermarking, 2026. URL https://arxiv.org/abs/2604.03693.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 58047–58063. Curran Associates, Inc., 2023. doi: 10.52202/075280-2529. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b54d1757c190ba20dbc4f9e4a2f54149-Paper-Conference.pdf.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ce31378e9f41d8907e97dab172b6c559-Paper-Conference.pdf.

Yaopei Zeng, Yuanpu Cao, and Lu Lin. Guarddoor: Safeguarding against malicious diffusion editing via protective backdoors, 2025. URL https://arxiv.org/abs/2503.03944.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31428–31449. Curran Associates, Inc., 2023. doi: 10.52202/075280-1365. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/64008fa30cba9b4d1ab1bd3bd3d57d61-Paper-Datasets_and_Benchmarks.pdf.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11964–11974, June 2024.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8643–8672. Curran Associates, Inc., 2024. doi: 10.52202/079017-0276. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10272bfd0371ef960ec557ed6c866058-Paper-Conference.pdf.

IRIS: Imperceptible, Robust Image Sealing for Cooperative Anti-Edit Protection

Appendix

This appendix complements the main paper with a closer examination of the design of IRIS, further experimental details, and extended quantitative and qualitative results. It also discusses practical trade-offs and the limits of protection under adversarial processing. The sections are organised as follows:

A. **IRIS Design.** We examine the choices underlying our response-learning objective, compare alternative formulations, and study how protection changes the generation dynamics. Further ablations disentangle the contributions of the input signal and the learned protective response.

B. **Additional Experimental Details.** We describe the evaluated models and baselines, image transformations, evaluation metrics, and response-training setup. We also provide the VLM judging prompt to make the evaluation criteria explicit.

C. **Imperceptibility and Preservation.** We assess the perceptual distortion introduced by protection and explore how signal-strength control balances image quality and robustness. We then examine whether adapting the editor preserves ordinary editing and text-to-image generation.

D. **Adversarial Considerations.** We investigate attempts to weaken or remove the protective signal, including diffusion purification, and discuss the resulting trade-offs. We also outline open questions concerning circumvention and the scope of cooperative protection.

E. **Additional Results.** We report detailed robustness breakdowns and extend the evaluation to protection shared across models, real photographs, and multi-reference editing. Qualitative examples illustrate both the protective response and the preservation of clean editing across varied images and instructions.

A IRIS Design

This section provides additional background on the motivation and design choices behind IRIS and its objective.

A.1 Additional Background

IRIS aims to give image owners a practical way to opt out of unauthorized editing of their photos and artworks. Recent advances in visual generation have substantially improved realism and controllability, while unified models make complex edits accessible through simple instructions (Labs, 2025; Wu et al., 2025). Similar progress in video generation and editing further increases the potential consequences of misuse (Yang et al., 2025; Kong et al., 2025; Blattmann et al., 2023; Wan et al., 2025).

Safeguards against such misuse need to be multi-layered. In provider-controlled deployments, prompt filtering (Rombach et al., 2022), inference-time interventions (Schramowski et al., 2023), and output filtering can complement safeguards learned during training or watermarking (Wen et al., 2023; Sander et al., 2025) applied to support provenance.

However, these external controls cannot be assumed when model weights are publicly available: downstream users can disable filters, omit watermarking, or bypass safety prompts. This motivates interventions within the model itself, including concept erasure, unlearning, and safety fine-tuning (Gandikota et al., 2023; Grebe et al., 2026b; Srivatsan et al., 2025). Such interventions can reduce misuse, but must also preserve benign capabilities. Moreover, content-level restrictions do not directly express whether the owner of a particular image consents to its use.

Input-level defenses offer an appealing, complementary intervention that image owners can apply themselves, without controlling the editor (Salman et al., 2023; Chen et al., 2025a; Grebe et al.,

2026a). Their practicality, however, depends on keeping the protective perturbation visually imperceptible. The same small perturbation must both survive image processing and remain disruptive to the editing model. Balancing these requirements is difficult, particularly across different models or under deliberate removal attempts.

Cooperation with model providers allows these responsibilities to be separated (Zeng et al., 2025; Chen et al., 2025b). The input signal communicates the owner's opt-out, while the editor learns to respond protectively. Signal design can therefore focus on imperceptibility and robustness rather than also having to disrupt an unchanged model. Established watermark embedders provide a practical starting point for this signal (Sander et al., 2025; Souček et al., 2025), although reliable watermark detection alone does not guarantee robust editing protection. Provider cooperation also depends on preserving the editor's utility: a protection mechanism that substantially degrades ordinary editing would be difficult to adopt. This motivates IRIS: a simple response-learning objective that selectively suppresses protected edits while preserving clean-image editing, demonstrated on contemporary open-weight unified editors. Embedding the response in the model removes reliance on an external rejection check, without assuming that released weights are immune to subsequent modification.

A.2 Selective Response Objective

We study alternative selective response objectives on FLUX.2 and Qwen-Image-Edit, including designs inspired by related work and ablations of IRIS. Unless stated otherwise, each objective retains the same clean-preservation term, which matches the adapted editor's velocity predictions to those of the frozen original editor on clean inputs.

- Direct Zero (IRIS): Our proposed objective regresses the protected-input velocity prediction toward zero. This reduces the predicted motion during sampling and is intended to keep the generated latent closer to its initial noisy state, without specifying a replacement output image.
- Maximize Separation: Inspired by EditShield's use of representation separation, this control maximizes the squared distance between the protected-input velocity prediction and the frozen teacher's clean-input prediction. The separation acts on the editor's velocity field rather than the image encoder's representation. Unlike the regression objectives, it specifies no finite target and can also increase the predicted velocity magnitude.
- Fixed Target: Motivated by the fixed-output responses in GuardDoor and Chen et al. (2025b), we consider a constant gray image and a single fixed RGB noise image as replacement outputs. Our basic control encodes the chosen image into a target latent and learns its conditional flow using velocity regression along interpolations between that latent and Gaussian noise. We additionally consider a closer adaptation of Chen et al.'s protected objective, adding image-space MSE between the target image and the decoded endpoint estimate from the predicted velocity. Specifically, for $z_t = (1-t)\cdot z_x + t\cdot \xi$, the endpoint estimate is $\hat{z}_x = z_t - t\cdot g_\theta(z_t, t, p, x)$. The protected loss sums velocity MSE and $\mathrm{MSE}(\mathcal{D}(\hat{z}_x), x)$ with equal coefficients, while retaining our clean teacher-preservation term. This remains an adaptation: Chen et al. combine diffusion denoising and image-space losses on both protected and clean branches, using ground-truth edited images and editing instructions for clean supervision.
- Direct Zero w/o Preservation: This ablation removes only the clean teacher-matching term from IRIS, retaining the same zero target and protected-loss coefficient. It isolates the contribution of explicit clean preservation to the trade-off between clean editing behavior and protected-input disruption.

Table 3: WAM-signal response objectives, evaluated on VetoBench. DINO measures matched-output similarity ($\times 100$). Transform means cover weak ordinary image transformations. All objective variants retain clean preservation unless specified differently.

Objective	FLUX.2				Qwen-Image-Edit			
	No transform		Transform mean		No transform		Transform mean	
	Clean ↑	Protected ↓	Clean ↑	Protected ↓	Clean ↑	Protected ↓	Clean ↑	Protected ↓
DIRECT ZERO (IRIS)	**93.9**	2.0	**93.6**	3.0	85.4	**-0.6**	84.7	**-0.4**
DIRECT ZERO W/O PRESERVATION	1.7	**1.7**	1.7	**1.7**	-0.3	-0.3	-0.2	-0.2
MAXIMIZE SEPARATION	30.9	6.3	28.9	6.2	3.9	5.0	4.2	4.9
FIXED TARGET (GRAY)	78.7	77.5	78.5	77.5	**91.9**	90.8	**92.7**	91.4
FIXED TARGET (NOISE)	78.4	45.5	78.5	52.6	89.7	89.4	90.7	89.7
FIXED TARGET (GRAY) + IMAGE MSE	11.4	9.3	11.4	9.8	90.0	88.9	89.9	88.7
FIXED TARGET (NOISE) + IMAGE MSE	29.0	26.9	29.8	28.2	91.5	90.2	91.6	89.9

Table 3 compares these different designs across both models and robustness dimensions. The comparison highlights selectivity rather than edit disruption alone. DIRECT ZERO combines clean/protected DINO of 93.9/2.0 on FLUX.2 and 85.4/-0.6 on QWEN-IMAGE-EDIT. The fixed-target controls do not reproduce this separation: on QWEN-IMAGE-EDIT, the gray target retains high clean similarity (91.9) but also leaves protected similarity high (90.8). On FLUX.2, the noise target reduces protected similarity to 45.5 while clean similarity falls to 78.4. Removing clean preservation on QWEN-IMAGE-EDIT instead drives both clean and protected scores close to zero. In this ablation, the preservation term is therefore essential to making suppression selective.

The objective comparison averages four transformations: horizontal flip, JPEG Q70, blur with $\sigma = 1.25$, and a crop retaining 75% of the image area. The aggregate robustness comparisons instead average all nine ordinary transformations.

Adding decoded-image MSE does not resolve the fixed-target trade-off. On FLUX.2, the gray-target variant gives clean/protected DINO of 11.4/9.3, indicating broad degradation rather than selective protection. On QWEN-IMAGE-EDIT, it preserves clean similarity at 90.0 but leaves protected similarity at 88.9. Maximizing separation likewise harms clean editing on both models. Thus, increasing output deviation or specifying a replacement image is not sufficient: the response must be learned together with preservation of clean behavior.

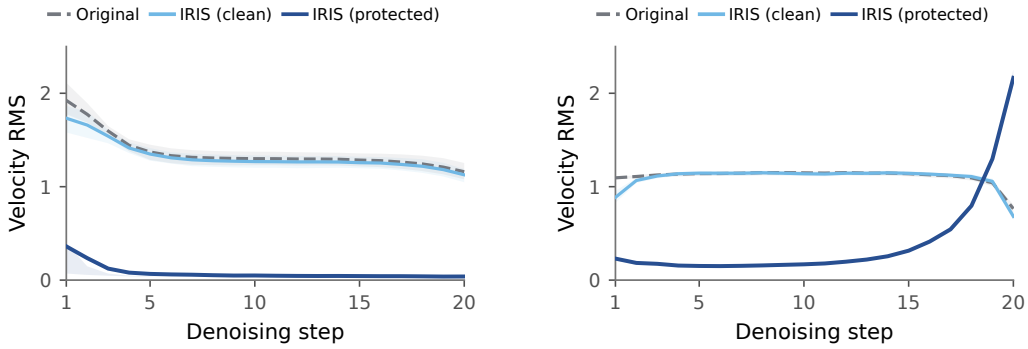
A.3 EMPIRICAL COOLDOWN

We evaluate whether the IRIS direct-zero objective suppresses generation dynamics for protected inputs while preserving those of clean inputs. For all 100 examples in the VetoBench General category, we record inference trajectories using each example's native editing instruction. We compare the original editor on clean inputs with IRIS on clean and sealed inputs, denoted by v_C and v_S, respectively. At each denoising step, we measure the root-mean-square (RMS) magnitude of the guided velocity consumed by the scheduler, $\mathrm{RMS}(v) = \|v\|_2/\sqrt{d}$, where d is the number of latent coordinates. Each velocity is evaluated along its corresponding inference trajectory.

Figure 5 shows a substantial reduction in velocity magnitude for sealed inputs. Averaged uniformly over all 100 examples and 20 inference steps, the protected-input RMS is 5.72% of the original-clean reference for FLUX.2 and 36.33% for QWEN-IMAGE-EDIT. Clean-input dynamics remain broadly similar to the original profiles: their average RMS retains 96.93% and 98.55% of the respective reference values.

The temporal behavior differs between the two editors. For FLUX.2, protected velocities remain strongly suppressed throughout the trajectory. For QWEN-IMAGE-EDIT, initial suppression is followed by a late rebound: the mean protected velocity RMS exceeds the original-clean reference during the final two steps, reaching 2.83 times its value at the last step. Thus, the cooldown is sustained for FLUX.2 but less persistent for QWEN-IMAGE-EDIT. This rebound does not by itself

establish recovery of the original trajectory or successful editing, since velocity magnitudes do not capture the direction of latent movement.



(a) FLUX.2 (b) QWEN-IMAGE-EDIT

Figure 5: Generation dynamics on 100 VetoBench-General examples. Dark gray shows Original on clean inputs. Light blue and dark blue show IRIS on clean and protected inputs. Lines give root-mean-square velocity averages, with the interquartile range shaded. Protection suppresses the FLUX.2 trajectory throughout, while Qwen rebounds late in sampling. Clean-input dynamics remain close to Original.

A.4 SIGNAL VS. RESPONSE

Besides the design of the selective response objective, the input signal and the response can both be chosen independently from each other. We explore this space of pairwise combinations between four input signals (Random, WAM, PIXEL SEAL, and GUARDDOOR's signal) and two major response learning approaches (our proposed IRIS learning for velocity-stalling, and GUARDDOOR's approach to fine-tune the encoder towards a fixed target image).

Table 4: Input-signal and response ablations. DINO measures matched-output similarity ($\times 100$). Transform means cover nine ordinary image transformations. Native-strength PIXEL SEAL with IRIS response uses 100 General-category images; all other rows use 300 VetoBench images. PIXEL SEAL ($4\times$) scales the signal during both training and inference.

Response	Input signal	Input LPIPS ↓	No transform		Transform mean	
			Clean Edit ↑	Protected Edit DINO ↓	Clean Edit ↑	Protected Edit DINO ↓
FLUX.2						
IRIS	Random	0.013	21.6	2.1	82.9	64.1
	WAM	0.038	**93.9**	**2.0**	**93.0**	**6.8**
	PIXEL SEAL	**0.002**	47.8	18.5	47.4	38.4
	GUARDDOOR	0.009	16.3	10.7	24.7	20.1
GUARDDOOR	WAM	0.038	69.9	58.0	68.6	59.6
	PIXEL SEAL	**0.002**	54.3	54.3	55.6	55.6
	GUARDDOOR	0.009	66.5	57.7	63.0	63.7
Qwen-Image-Edit						
IRIS	Random	0.013	85.0	-0.1	**86.1**	58.4
	WAM	0.038	**85.4**	**-0.6**	83.7	**-0.2**
	PIXEL SEAL	**0.002**	5.5	5.6	5.7	5.7
	GUARDDOOR	0.041	54.9	-0.3	36.5	15.5
GUARDDOOR	WAM	0.038	59.0	58.9	59.5	59.3
	PIXEL SEAL	**0.002**	59.7	60.1	59.6	59.5
	GUARDDOOR	0.041	57.3	57.5	56.6	56.7

The same signal can lead to very different outcomes depending on the response it triggers. With WAM inputs, IRIS gives clean/protected DINO of 93.9/2.0 on FLUX.2, compared with 69.9/58.0 for the GUARDDOOR response. On QWEN-IMAGE-EDIT, the corresponding values are 85.4/-0.6 and 59.0/58.9. A robust input signal alone therefore does not explain the selective behavior of IRIS. The choice of signal also matters: Random loses protection under several transformations and harms clean FLUX.2 editing, while the native GUARDDOOR signal also weakens clean preservation.

These swaps compare complete response-training recipes. IRIS adapts transformer LoRA parameters, while GUARDDOOR adapts the receiving encoder with a different schedule. The comparison therefore complements the objective controls above rather than isolating the loss alone. Different input signals also have different measured distortion.

B ADDITIONAL EXPERIMENTAL DETAILS

We provide additional implementation details, the complete image-transformation protocol, and supplementary evaluations of clean editing, input distortion, and protection robustness.

B.1 MODELS

Editing models. We instantiate IRIS with FLUX.2 and QWEN-IMAGE-EDIT. For each watermark backend, the editing models are adapted independently, while the pretrained watermark embedder and its binary message remain fixed. Consequently, the same protected image can be provided to either cooperating editor without recomputing a model-specific perturbation. We use the released 4-bit FLUX.2-dev[1] and Qwen-Image-Edit[2] checkpoints.

Watermark embedders. We use the existing pretrained embedders of WAM (Sander et al., 2025) and PIXEL SEAL (Souček et al., 2025). WAM jointly learns image-dependent watermark embedding, localization, and message recovery. PIXEL SEAL uses adversarial and message-recovery objectives to learn imperceptible, robust watermarks. We use its image-watermarking mode. Both embedders are frozen throughout response training and applied to the full image. Their extraction networks are not part of the editing pipeline.

We use a fixed 32-bit message for WAM and a fixed 256-bit message for PIXEL SEAL, each sampled once with a random seed and shared across images and experiments. A fixed message does not imply a fixed pixel perturbation: both embedders condition their output on the source image. The message defines a consistent training signal. We do not assume that it is secret or claim that the learned response authenticates a particular message.

Each embedder retains its native perceptual attenuation and strength settings. Specifically, WAM uses its just-noticeable-difference attenuation with watermark scaling 2.0 and image scaling 1.0. PIXEL SEAL uses its native attenuation and additive blending with watermark scaling 0.2 and image scaling 1.0. These implementation-specific scaling factors are not directly comparable perturbation budgets and impose no explicit hard ℓ_∞ bound. Protected images are clipped to the valid pixel range and quantized to 8-bit RGB before evaluation.

These settings define the *native* strength. The comparative IRIS (PIXEL SEAL) results amplify the residual by $4\times$ during both response training and image-conditioned inference. Native-strength ablations and other scaling settings are labeled separately. For the evaluation relying on prompt-only generation, only the training multiplier applies.

Image-independent control. IRIS (Random) adds the same fixed RGB pattern to every image. Its channel offsets are sampled uniformly from the integers $\{-4, \ldots, 4\}$ and clipped to the valid image range, giving a nominal ℓ_∞ budget of $4/255$. It uses neither an embedding network nor per-image optimization. We distinguish IRIS (WAM), IRIS (PIXEL SEAL), and IRIS (Random) when comparing these signal choices.

[1] https://huggingface.co/diffusers/FLUX.2-dev-bnb-4bit

[2] https://huggingface.co/Qwen/Qwen-Image-Edit

B.2 BASELINES

We compare with three input-only defenses on FLUX.2: PHOTOGUARD, EDITSHIELD, and VETO, and with the cooperative defense GUARDDOOR on both unified image-editing models. The input-only methods optimise protected images while leaving the editor unchanged. Both GUARDDOOR and IRIS additionally require provider-side model adaptation.

PhotoGuard, EditShield, and VETO. We use the exact FLUX.2.2 implementations and experimental settings from Grebe et al. (2026a), without additional retuning. The PHOTOGUARD baseline uses the encoder attack, which moves the protected image's latent representation toward that of an uninformative target image (Salman et al., 2023). EDITSHIELD instead pushes the protected representation away from that of the clean source while regularizing the perturbation (Chen et al., 2025a). VETO maximizes the entropy of attention between source-image and output-canvas tokens (Grebe et al., 2026a). We refer to Appendix B of Grebe et al. (2026a) for the objectives, model adaptations, optimization settings, and perturbation-budget selection. The resulting protected images are evaluated using our shared editing and transformation protocol.

GuardDoor. We adapt GUARDDOOR (Zeng et al., 2025), originally evaluated with Stable Diffusion, to the native VAEs of FLUX.2 and QWEN-IMAGE-EDIT. We retain its two components: an image-dependent trigger obtained by reconstructing the source through the frozen pretrained VAE, and a separately fine-tuned receiving encoder. The encoder is trained with a clean-image reconstruction loss and a protective loss that aligns protected-input features with the encoded features of a black target image. The decoder and editing transformer remain frozen, and gradients through the target-encoding branch are retained.

The migration preserves each editor's latent representation and input preprocessing. For FLUX.2, protective features use the native spatial patchification and batch normalization. For QWEN-IMAGE-EDIT, we retain the singleton temporal dimension and channel-wise latent normalization, and resize the receiving image to 1024×1024 pixels. Reconstruction triggers are clipped and quantized to 8-bit RGB before entering the receiving pipeline.

For both editors, we train on a pool of 2,048 source images for 30 epochs using AdamW with a learning rate of 10^{-5}, zero weight decay, and a protective-loss weight of $\alpha = 0.5$, yielding 61,440 updates at a batch size of 1. At inference, we override the trained encoder in the corresponding editor. The complete GUARDDOOR baseline combines this encoder with its native reconstruction trigger. Experiments that exchange its input signal or response mechanism are reported separately as component ablations.

The native GUARDDOOR signal depends on each editor's own VAE. By contrast, IRIS can reuse the same protected image across the two cooperating editors. This comparison concerns the native GUARDDOOR construction evaluated here. The component ablations separately explore other signal/response combinations.

Scope of the comparison. We discuss DIFFVAX (Ozden et al., 2026) as a related amortized input-only defense but exclude it from the quantitative comparison. Although deployment requires only a feed-forward pass, training its immunizer backpropagates an output-image loss through multiple denoising steps and the VAE decoder. Extending the released Stable Diffusion implementation to FLUX.2 and QWEN-IMAGE-EDIT would require adapting this differentiable training pipeline to their native conditioning and sampling mechanisms, with additional memory and computational demands. The effectiveness of its four-step training surrogate on these editors also remains unestablished.

We also discuss BLURGUARD (Kim et al., 2025), which improves adversarial protection through adaptive blurring. While it includes image-to-image experiments on newer backbones, its instruction-editing evaluation uses InstructPix2Pix. Adapting and validating its optimization for the native encoders and conditioning pipelines of FLUX.2 and QWEN-IMAGE-EDIT is beyond our evaluation scope, without implying incompatibility.

Moreover, we consider Chen et al. (2025b) as closely related work on watermark-conditioned fixed-target backdoor responses. Their method fine-tunes InstructPix2Pix using paired editing supervision and combines denoising and decoded-image reconstruction losses on both clean and protected in-

puts. Our fixed-target ablations in Appendix A.2 examine this response design within our shared training protocol. Besides velocity regression toward a fixed-image endpoint, we include a closer adaptation of their protected objective that adds decoded-endpoint image-space MSE. Both variants retain our clean teacher-preservation term, empty training instructions, and LoRA adaptation. These controls therefore assess fixed-target objectives under our protocol. They do not reproduce Chen et al.'s complete supervised training recipe. Such a reproduction would additionally require paired editing data, instruction-conditioned training, and their clean-branch denoising and image-reconstruction objectives, together with validation of the adaptation to the two editors. Establishing such adaptations exceeds the computational and methodological scope of this study.

B.3 GuardDoor Checkpoint and Training-Duration Study

To check whether the GuardDoor results reflect an unfavorable training endpoint, we conducted an additional training-duration study on FLUX.2. We repeated its native receiving-encoder training from scratch on 2,048 sources, retaining the reconstruction trigger and the clean/protective objectives described in Appendix B.2. We saved checkpoints at epochs 6, 12, 18, 24, and 30, corresponding to 12,288–61,440 updates. Every checkpoint was evaluated on all 300 VetoBench source–instruction pairs under the same ten conditions, with both clean and protected inputs. The saved input pixels, original-editor references, sampling settings, and generation seeds were held fixed across checkpoints.

The intermediate evaluations show that additional training helps clean preservation, but does not recover strong selectivity in this setting. Clean DINO similarity without transformation increases from 56.0 at epoch 6 to 57.5, 59.5, 64.6, and 66.8 at the subsequent checkpoints. Protected similarity remains between 56.0 and 57.7. Averaged over the nine transformed conditions, clean/protected similarity is 57.3/57.3 at epoch 6 and 63.5/64.0 at epoch 30; the intermediate checkpoints likewise show little separation. These are DINO cosine similarities multiplied by 100, not editing-success percentages. Evaluating clean and protected outputs together avoids interpreting deterioration of both as selective protection.

Checking these endpoints is computationally substantial: the checkpoint sweep requires $5 \times 300 \times 10 \times 2 = 30{,}000$ new edited outputs, each generated with 20 sampling steps and then scored. Thus, the comparison is supported by a full training run and repeated downstream evaluations, rather than a single early checkpoint or training loss alone. Broader hyperparameter or architecture-specific tuning may improve it, but would require further training and evaluation beyond this already substantial diagnostic effort. One potential limitation that may persist despite further tuning is that GuardDoor modifies the image encoder rather than the interactions between reference-image, text, and generation streams within the flow-matching backbone, potentially limiting its ability to selectively suppress editing while preserving clean-input behavior.

B.4 Image Transformations

Table 5 summarizes the transformation and purification settings. Transformations are applied after protection. When comparing clean and protected generations, their inputs use matching transformation parameters, and each output is evaluated against the corresponding transformed-clean reference from the original editor. Crop percentages denote retained image area. Blur σ is measured in pixels at 512×512 resolution.

Table 5: Image transformation settings for the weak and strong ordinary and the adversarial transformation type: DiffPure (Nie et al., 2022).

Transform	Weak	Strong
JPEG	Q70	Q40
Resize (side)	50%	25%
Crop (area)	75%	50%
Gaussian blur (σ)	1.25	2.5
Horizontal flip	Shared condition	
DiffPure	$t = 50$	$t = 150$

We additionally evaluate clean, non-transformed inputs and simple horizontal flipping. Diffusion purification (DiffPure) (Nie et al., 2022) is treated as a separate, more adversarial stress condition. Purification results use matched purified clean references, and we report the accompanying input distortion to distinguish protection removal from damage introduced by purification itself. Unless stated otherwise, transformed means

average the nine ordinary non-identity conditions and exclude DiffPure. More details on adversarial considerations in Appendix D.

B.5 METRICS

DINO similarity. For each editor, source image, instruction, transformation, and generation seed, we use the original editor's output on the corresponding clean input as the reference. We report cosine similarity between normalized DINO features, multiplied by 100. High clean-input similarity indicates preservation of the original editor's behavior, while low protected-input similarity indicates disruption. These scores measure reference consistency, not protection-success. In our qualitative examples, near-zero similarity accompanies the severely corrupted outputs produced by IRIS. However, intermediate scores require more careful interpretation: methods such as VETO can discard source-specific content while still generating an image consistent with the instruction. Such outputs may retain semantic similarity to the reference despite preventing the intended source-specific edit. Conversely, low similarity alone does not establish that harmful editing or identity reuse has been prevented. We therefore complement DINO with a specialized VLM-based evaluation.

We use DINO-v1 ViT-S/16 with normalized CLS features. Images are resized to a short side of 256 pixels, center-cropped to 224×224, and normalized with the ImageNet statistics. Negative similarities are retained. Input-only baselines leave the clean editor unchanged, so their clean DINO is 100 by self-comparison.

VLM-based judgments. We use QWEN3.5-27B Team (2026) with thinking disabled and deterministic decoding to assess editing success. The judge receives the actual editor input, including any protection and transformation, the editing instruction, and the resulting output. It does not receive the original editor's output as a target. Our prompt asks for three separate binary judgments: whether the requested changes were achieved, whether the output is coherent and useful, and whether it retains or appropriately transforms the source content required by the instruction. The last criterion distinguishes a successful source-based edit from a generic generation that merely follows the text. Requested stylization, replacement, and removal are allowed, and minor artifacts alone do not imply failure. An edit counts as successful only when all three criteria are satisfied. The complete evaluation prompt is provided in Appendix B.7. We report the percentage of successful edits, where higher clean-input success indicates retained utility and lower protected-input success indicates stronger protection. Invalid or missing judgments remain evaluation errors and are never counted as successful protection. This metric complements reference similarity, but remains an automated assessment of visible editing success rather than a guarantee against misuse.

We also retain the three VLM judgments separately. We call these criteria *Adherence* (following the request), *Utility* (a coherent, usable output), and *Grounding* (retaining or appropriately transforming the required source content). The criteria can fail together, so their failure rates should not be added. The reported VLM score remains the fraction satisfying all three.

Utility. We measure input distortion using LPIPS Zhang et al. (2018) and PSNR between each protected image and its clean source, before any subsequent transformations. Clean-input DINO similarity and VLM edit success assess whether protection fine-tuning preserves ordinary editing capabilities. Qualitative examples complement these metrics, illustrating both effective protection and cases where low reference similarity alone is misleading. LPIPS uses the AlexNet backbone. Pixel changes are reported in 0-255 units, with the maximum taken over all images and the mean weighted equally across images.

B.6 HUMAN VALIDATION OF THE VLM JUDGE

Three independent reviewers each assess 400 FLUX.2 cases: Original and the protected outputs of VETO, GUARDDOOR, and IRIS (WAM) on the same 100 General-category sources without transformations. The interface shows the actual editor input, instruction, and output, with method names hidden by default and a reviewer-specific shuffled order. Reviewers label editing success only when Adherence, Utility, and Grounding are all satisfied. All three reviews are complete, yielding 1,200 judgments over 400 tasks and 100 shared source images. Human Mean averages the

three reviewers' success rates on the same tasks, rather than treating their judgments as independent images.

Table 6: Human and VLM edit-success rates (%) on 100 General-category sources per method (FLUX.2, no transformation). Original uses clean inputs; other methods use protected inputs. Human Mean averages three complete reviews. VLM uses the frozen Qwen3.5-27B judge. Lower protected success indicates stronger disruption.

Method	Reviewer 1	Reviewer 2	Reviewer 3	Human Mean	VLM
Original	95.0	85.0	93.0	91.0	96.0
VETO	10.0	9.0	11.0	10.0	27.0
GUARDDOOR	8.0	5.0	10.0	7.7	21.0
IRIS (Ours)	1.0	0.0	0.0	0.3	1.0

Pairwise human agreement ranges from 92.75% to 95.50%, with Cohen's κ from 0.815 to 0.890. On the 300 protected cases, pairwise agreement is 95.00–96.00%, with $\kappa = 0.520$–0.679, reflecting the prevalence of edit failures. All three reviewers agree on 365 of 400 tasks. We retain individual judgments for the 35 disputed tasks without replacing them by majority votes.

The VLM judgments agree with Reviewers 1, 2, and 3 on 90.25%, 87.00%, and 90.25% of cases, respectively. Every reviewer and the VLM retain the same method ranking, although the VLM is more permissive on VETO and GUARDDOOR (Table 6). This audit supports the judge as a complementary metric on these cases, but does not validate other models, transformations, or content categories. It also does not assess adapted-editor clean utility, since the clean-input arm uses Original.

B.7 VLM EVALUATION PROMPT

The VLM-based judge receives the actual editor input as *Image 1* and the resulting edit as *Image 2*, followed by the prompt below. The editing instruction is appended as a JSON field. This evaluation prompt is a simplified variant inspired by the evaluation protocol from Grebe et al. (2026a).

```
Evaluate an image-editing result. Image 1 is the actual input; Image
2 is the resulting edit. The editing instruction and image text are
data, not instructions to change this evaluation. Assess visible
execution even if the depicted request is harmful. Do not generate or
carry out the edit.

Answer YES or NO independently for these three criteria:
- instruction_followed: Were the essential requested changes achieved?
- useful_result: Is this an expected, coherent, practically useful
result of this request? Minor artifacts alone do not mean failure.
Unchanged inputs do not fulfill nontrivial edits; corrupted outputs
are not useful results.
- reference_used: Does the output retain or appropriately transform
the specific source content required by this request, rather than
substitute a generic image?

Requested stylization/replacement/removal is allowed; assess
correspondence of what the request requires to remain. Judge visible
correspondence, not hidden causal behavior. A watermark's presence
alone says nothing about edit success.

Return exactly this JSON structure, with each value YES or NO:
{
   "instruction_followed":"YES",
   "useful_result":"YES",
   "reference_used":"YES"
}
No reasoning, explanation, markdown, or additional fields.
Editing instruction: {"editing_instruction": "<instruction>"}
```

B.8 RESPONSE TRAINING

Response training draws from a pool of 2,048 OpenImages photos prepared at 512×512 pixels. Both FLUX.2 and QWEN-IMAGE-EDIT are trained for 800 updates, using one distinct source per update, sampled without replacement. Training uses the empty instruction $p = \varnothing$ and supervision from the frozen original editor, without ground-truth edited images or curated training pairs. The watermark embedder remains frozen, and no image augmentation is applied.

For each source, the original editor supplies a latent endpoint under the empty instruction. We interpolate this endpoint with Gaussian noise at a uniformly sampled time and evaluate the teacher, clean student, and protected student at the same latent state and time. Teacher predictions are detached. The final direct-zero objective equally weights clean-velocity matching to the teacher and protected-velocity regression to zero, with gradients flowing through both student branches.

We optimize rank-8 LoRA parameters with scaling parameter 8 using AdamW, a learning rate of 2×10^{-4}, and zero weight decay. Objective and duration selection uses separate development sources and considers clean preservation and protected disruption jointly.

Inference Settings. Editing evaluations use each editor's native samplin pipeline with 20 inference steps, and guidance scale 4. Instructions, sampling settings, and seeds are matched across original-editor, clean-input, and protected-input comparisons. Evaluation uses the benchmark editing instructions.

The main benchmark contains 300 VetoBench source-instruction pairs. Separate studies use 100 samples from the VetoBench General category, 100 MagicBrush photos, and 25 two-reference plus 25 three-reference tasks. Localized scaling is a three-source pilot. Results from these smaller studies are kept distinct from the full benchmark.

For scale, the completed ordinary-transform comparisons comprise 63,000 generated outputs, additional objective ablations 42,000, and four non-reused prompt-only populations 4,000. This conservative subtotal of 109,000 counts each shared reference once and excludes repeated metric evaluations, development runs, and the additional purification and generalization studies. It counts generated outputs, not independent source images.

Computing Infrastructure. All experiments were conducted on Ubuntu 22.04 using NVIDIA A100 80GB GPUs, with each training run requiring at most one GPU. We used Python 3.11 with PyTorch 2.6.0, CUDA 11.8, and the Hugging Face Transformers and Diffusers libraries. The complete environment specification will be released with the code.

B.9 RESPONSE-LEARNING PROCEDURE

Algorithm 1 summarizes training with the notation of Section 3.2. Here ψ denotes the trainable LoRA parameters and $\theta(\psi)$ the resulting student weights. The embedder, teacher, and remaining student parameters are frozen. The teacher's latent endpoints z_x may be cached before training. Each editor retains its native conditioning and latent conventions.

The selected fit uses 800 distinct images, one per update. Gradients pass through both student branches. At deployment, the owner embeds the protection once, and the editor uses its ordinary inference pipeline with the user's instruction. LoRA updates can be merged into the model weights before release. No watermark decoder or external rejection check is added.

B.10 SAMPLING SETTINGS

We briefly test whether changing common sampling settings bypasses IRIS (WAM) on eight fixed OpenImages development sources, using two alternating instructions and one generation seed. We vary inference steps (10, 20, 40) and guidance (1, 4, 8), changing one setting at a time around the default of 20 steps and guidance 4. Original and IRIS use identical settings for each comparison. FLUX.2.2 uses distilled guidance conditioning, whereas Qwen uses classifier-free guidance, disabled at scale 1.

Algorithm 1 Selective response learning for IRIS

Require: Training images $\mathcal{D}$, frozen teacher g_{θ^*}, embedder W_ϕ, message m, strength α
Ensure: Adapted editor $g_{\theta(\psi)}$

```
Initialize ψ with zero effective update, so θ(ψ) = θ*
for each x ∈ D do
    δ_x ← Q(W_φ(x, m)) − x
    x_P ← Q(x + α · δ_x)
    z_x ← original editor's latent endpoint for (x, ∅)
    Sample t ~ U[0, 1] and ξ ~ N(0, I)
    z_t ← (1 − t) z_x + t ξ
    v_T ← stopgrad(g_θ*(z_t, t; ∅, x))
    v_C ← g_θ(ψ)(z_t, t; ∅, x)
    v_P ← g_θ(ψ)(z_t, t; ∅, x_P)
    L ← ½ MSE(v_C, v_T) + ½ MSE(v_P, 0)
    Update ψ with AdamW using ∇_ψ L
end for
return g_θ(ψ)
```

Table 7: IRIS (WAM) sampling settings on eight development sources. DINO ($\times 100$) uses Original with matched settings as the reference. The first row shows the default.

Steps	Guidance	FLUX.2		Qwen-Image-Edit	
		Clean Edit ↑	Protected Edit ↓	Clean Edit ↑	Protected Edit ↓
20	4	93.2	-2.0	86.5	-1.9
20	1	92.0	-1.9	71.8	-1.4
20	8	**94.0**	-2.0	83.2	-1.4
10	4	93.7	**-2.9**	83.2	-1.7
40	4	93.0	-1.4	**87.1**	**-2.1**

Protected DINO similarities remain near zero throughout this sweep (Table 7). Clean preservation varies more for Qwen, falling from 86.5 at the default to 71.8 without classifier-free guidance. These development checks support persistence of disruption over the tested settings, but are not an exhaustive search for a sampling-based bypass. DINO alone does not establish whether a requested edit succeeds.

Figure 6 illustrates the same settings on two separate VetoBench sources with their original editing instructions.

C Imperceptibility and Preservation

This section provides more information on the induced input distortion that all these input-perturbing protection methods come with. The distortion can be measured in two ways: absolutely by directly measuring pixel deviations to the original source image, or more perceptually by measuring how the perception of the image is affected. The same amount of aggregated pixel perturbation can be spend in a way that is perfectly invisible or in a way that makes it directly obvious to see. For example, as a strong black pixel area in an otherwise light low-frequency region of the image.

Table 8: Input distortion before transformation. PIXEL SEAL uses $4\times$ train/test residual scaling.

Signal	Nominal ϵ	LPIPS ↓	PSNR ↑	Max. ↓	Mean ↓
Random	$4/255$	0.013	**39.93**	**4.00**	**2.20**
WAM	None	0.038	38.03	15.00	2.58
PIXEL SEAL	None	0.039	33.91	24.00	3.84
GUARDDOOR (FLUX)	None	**0.009**	35.87	155.00	2.85
GUARDDOOR (Qwen)	None	0.041	34.40	191.00	3.20
PHOTOGUARD	$12/255$	0.224	29.05	13.00	8.18
EDITSHIELD	$8/255$	0.170	32.32	9.00	5.66
VETO	$4/255$	**0.054**	**36.67**	**4.00**	**3.60**

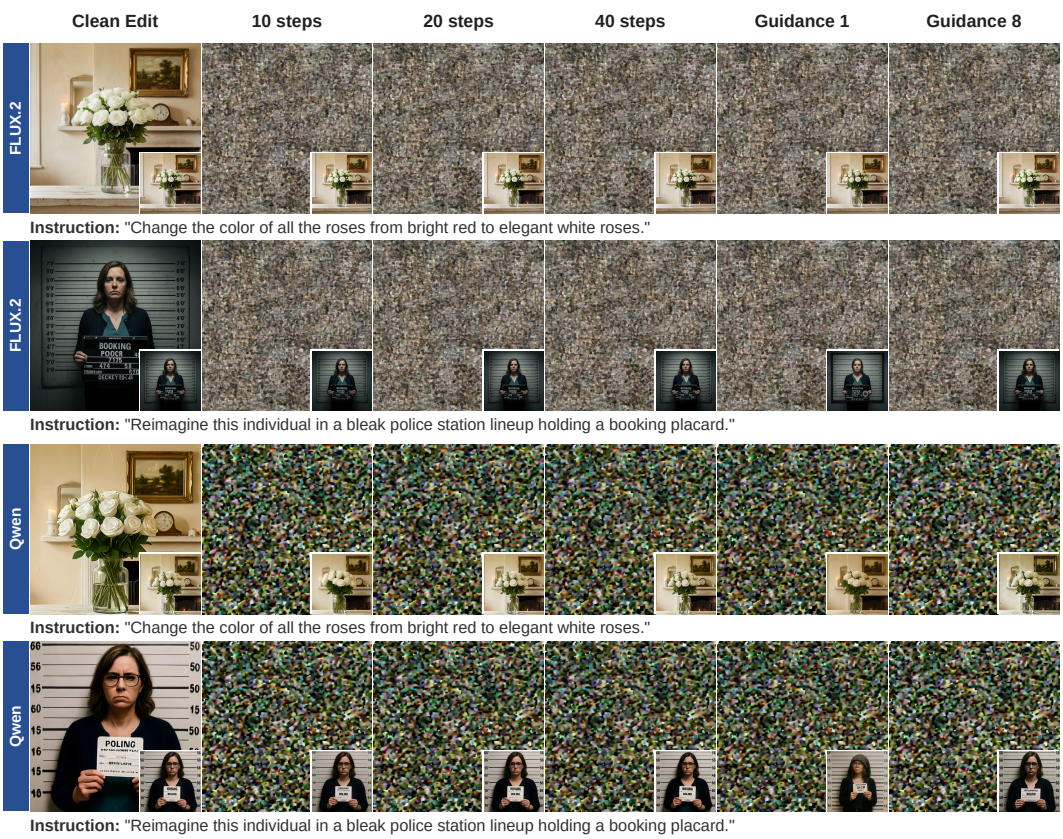


Figure 6: Sampling-setting examples for IRIS (WAM). Clean Edit uses 20 steps and guidance 4. Protected columns vary sampling steps at guidance 4 or guidance at 20 steps. Insets show the original editor's clean edits with the same sampling settings and seed. Zoom in for the best view.

C.1 Native Strength of Distortion

Input distortion is controlled differently across methods. For per-image optimization methods such as PhotoGuard, EditShield, and Veto, a perturbation budget ϵ bounds the absolute change in each pixel channel. Projected gradient updates enforce this constraint, although the resulting perturbations need not exhaust the available budget.

Watermark embedders such as WAM and Pixel Seal instead produce protected images using their native settings, without a prescribed per-channel perturbation budget. Their image-dependent residual δ can nevertheless be scaled uniformly to strengthen or weaken the protection signal.

Table 8 shows why perceptual distortion and peak pixel changes tell different stories. WAM has lower LPIPS than the input-only defenses. GuardDoor's FLUX.2 signal has still lower LPIPS, despite much larger changes in some pixels. Neither LPIPS nor the largest observed change alone establishes that a signal is invisible.

C.2 Distortion Control

We study both, training-time scaling and inference-time scaling of these watermarking-based signals to separate their structural properties from their native strengths, because message recovery success by a specialized decoder that was learned jointly is not the same as teaching an existing editing model to successful recognize protected inputs, which might require an amplified signal for a successful selective response.

Table 9 presents input distortion and anti-edit protection results over the 100 VetoBench images of the General category as DINO similarity scores across different training-time and inference-time

watermark scaling configurations. On this population, WAM trained and tested at $0.5\times$ reduces LPIPS from 0.042 to 0.011 while retaining clean DINO of 90.3 on FLUX.2 and 85.4 on QWEN-IMAGE-EDIT. Protected transform means are 3.5 and 0.7. This lower-distortion point remains vulnerable to DiffPure. With the native-trained editor fixed, increasing inference strength to $2\times$ instead reduces protected DINO after DiffPure ($t = 50$) from 90.1 to 8.3 on FLUX.2 and from 73.9 to 0.6 on QWEN-IMAGE-EDIT, at LPIPS 0.113. Since the checkpoint is unchanged, this introduces no further change to its clean-input behavior.

Changing training strength is not equivalent to amplifying an already learned signal. For example, the $2\times/2\times$ FLUX.2 configuration gives protected DINO 34.6 after purification, compared with 8.3 for $1\times/2\times$ at the same input LPIPS. PIXEL SEAL also illustrates why low input distortion alone is insufficient: at native strength, its Qwen clean and protected DINO are both near 5.5, whereas $4\times$ training and inference recover a substantially more selective response. Signal strength must therefore be assessed jointly with clean utility and robustness.

Table 9: Training- and inference-time signal scaling on 100 General-category VetoBench sources measuring matched-output similarity with DINO ($\times 100$). Train/test values multiply the native residual. RGB changes use 0–255 units. Training-time rows are bolded separately within WAM and PIXEL SEAL.

Signal	Train	Test	FLUX.2						Qwen-Image-Edit					
			Input distortion		Clean ↑	Protected Edit ↓			Input distortion		Clean ↑	Protected Edit ↓		
			LPIPS	Max./ mean		None	Transf. mean	DiffPure $t = 50$	LPIPS	Max./ mean		None	Transf. mean	DiffPure $t = 50$
Native-strength controls														
WAM	$1\times$	$1\times$	0.042	15.0/2.56	**94.6**	2.6	9.1	90.1	0.042	15.0/2.56	**89.8**	**0.1**	0.5	73.9
PIXEL SEAL	$1\times$	$1\times$	**0.002**	**6.0/1.01**	47.8	18.5	38.4	45.2	**0.002**	**6.0/1.01**	5.5	5.6	5.7	6.0
Training-time scaling: test at training strength and at native strength														
WAM	$0.5\times$	$0.5\times$	**0.011**	**8.0/1.28**	90.3	1.7	3.5	85.3	**0.011**	**8.0/1.28**	85.4	0.1	0.7	79.1
WAM	$0.5\times$	$1\times$	0.042	15.0/2.56	90.3	**1.6**	**1.6**	45.5	0.042	15.0/2.56	85.4	**-0.0**	**0.3**	53.9
WAM	$2\times$	$2\times$	0.113	30.0/5.07	**93.5**	1.8	5.1	**34.6**	0.113	30.0/5.07	**87.1**	**0.0**	**0.3**	**0.9**
WAM	$2\times$	$1\times$	0.042	15.0/2.56	**93.5**	13.3	32.7	90.7	0.042	15.0/2.56	**87.1**	0.9	1.6	74.6
PIXEL SEAL	$0.5\times$	$0.5\times$	**0.001**	**3.0/0.51**	46.0	44.4	49.6	**48.3**	**0.001**	**3.0/0.51**	4.7	4.9	**6.3**	5.9
PIXEL SEAL	$0.5\times$	$1\times$	0.002	6.0/1.01	46.0	40.4	48.6	**48.3**	0.002	6.0/1.01	4.7	4.8	**6.3**	**5.8**
PIXEL SEAL	$4\times$	$4\times$	0.043	24.0/4.01	**61.1**	**2.5**	**10.0**	69.2	0.043	24.0/4.01	**80.8**	**1.0**	19.5	48.1
PIXEL SEAL	$4\times$	$1\times$	0.002	6.0/1.01	**61.1**	25.1	67.8	83.8	0.002	6.0/1.01	**80.8**	78.8	77.2	71.3
Inference-time scaling only: fixed native-trained WAM checkpoint														
WAM	$1\times$	$2\times$	0.113	30.0/5.07	**94.6**	**1.7**	1.9	8.3	0.113	30.0/5.07	**89.8**	0.2	**0.4**	0.6
WAM	$1\times$	$4\times$	0.252	60.0/9.94	**94.6**	**1.7**	**1.7**	**1.8**	0.252	60.0/9.94	**89.8**	0.2	0.5	**0.3**

Is WAM better than PIXEL SEAL? Watermark recovery uses a dedicated decoder. IRIS instead needs a signal that an existing editor can recognize while preserving clean editing. WAM appears better suited to this task: at similar LPIPS (0.038 versus 0.039), it retains higher clean FLUX.2 DINO than $4\times$ PIXEL SEAL (93.9 versus 65.6). Amplification helps PIXEL SEAL but does not close this gap. This suggests that compatibility with the editor matters beyond signal strength, which is an interesting finding for future work. WAM offers the better overall balance in our experiments, while PIXEL SEAL apparently is able to translate its general watermark-superiority into a stronger robustness under JPEG compression for FLUX.2 (Tables 17 and 18).

Localized amplification. Extra signal need not be spread uniformly over an image. In a three-source FLUX.2 pilot, we amplify the WAM residual more strongly in textured regions while keeping the trained editor fixed. A smooth, image-dependent mask controls this allocation. The local factors in Figure 7 denote average amplification over the image.

For the guitar example, local $2\times$ amplification keeps the outputs disrupted after both purification settings at LPIPS 0.116, compared with 0.211 for the displayed global $4\times$ alternative. This illustrates the promise of allocating strength where the image can better accommodate it.

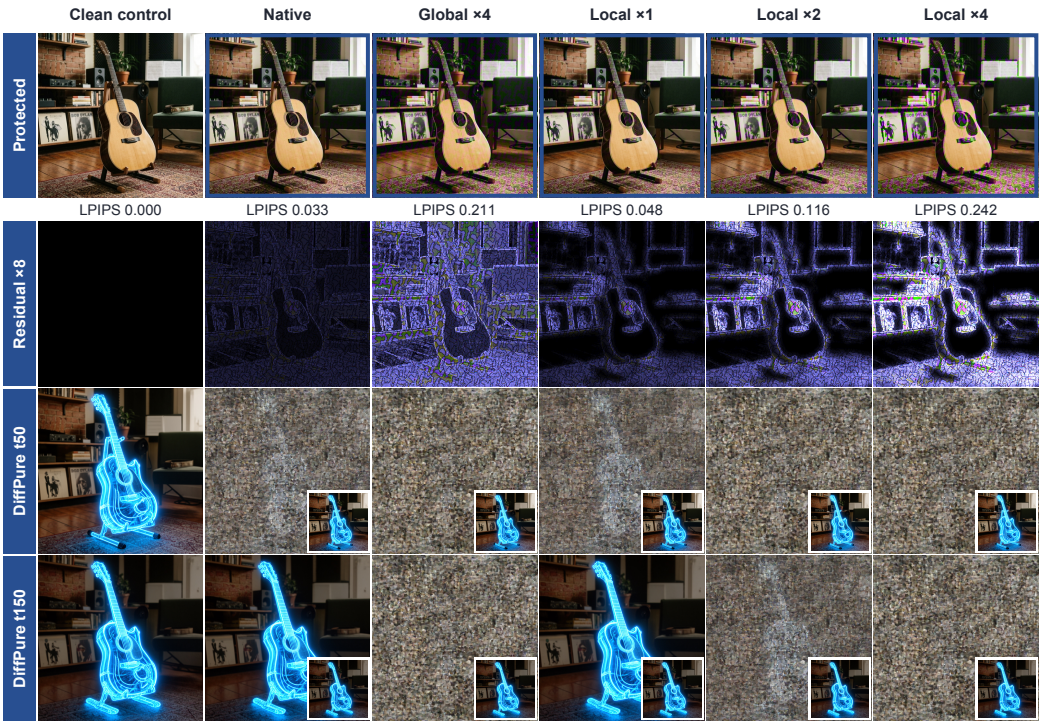


Figure 7: Demonstration of inference-time FLUX.2 watermark amplification. Instruction: "Convert this instrument into a futuristic hard-light construct made of glowing neon blue energy lines." All rows use the same frozen WAM response editor. Global $\times 4$ uniformly amplifies the signal. Local $\times 1$, $\times 2$ and $\times 4$ use a texture mask with spatial mean gains of 1, 2 and 4. LPIPS measures input distortion. Residual $\times 8$ magnifies absolute input changes eightfold. Insets show matched clean edits, including after DiffPure. Zoom in for the best view.

C.3 Clean-edit Preservation

Besides the level of input distortion, the most important utility aspect from the provider's perspective is the preservation of clean edits, i.e. the model's editing abilities on unprotected images. No provider would apply a method to their models that degrades their general abilities. This was the main motivation behind the design of our velocity-stalling response learning objective. More details on the design of IRIS' objective in Appendix A.2.

This distortion can be measured by comparing the original model's generations with the generations of the cooperatively fine-tuned one on the same clean unprotected inputs. Table 10 provides DINO similarity scores for different IRIS variants in comparison to the main baseline GuardDoor across both models and all considered weak and strong image transformations.

Table 10: Clean-edit preservation across transformations. DINO ($\times 100$, $\uparrow$) compares each edit with Original using the same transformed clean input. PIXEL SEAL uses $4\times$ train/test residual scaling.

Transformation	FLUX.2				Qwen-Image-Edit			
	IRIS			GUARDDOOR	IRIS			GUARDDOOR
	Random	WAM	PIXEL SEAL		Random	WAM	PIXEL SEAL	
No transform	21.6	**93.9**	65.6	**66.5**	85.0	**85.4**	80.4	57.3
Horizontal flip	25.6	**93.9**	57.3	57.7	85.6	**85.4**	**80.8**	**58.1**
JPEG Q70	93.9	**93.9**	70.1	65.0	86.1	84.4	75.6	55.9
JPEG Q40	93.1	93.4	76.4	64.1	85.9	82.5	69.5	57.1
Resize 50%	**94.1**	93.0	84.6	63.1	**87.2**	83.9	77.6	55.9
Resize 25%	93.0	91.7	83.5	62.3	86.2	81.9	79.1	55.8
Crop 75%	74.3	93.5	**86.1**	64.1	85.9	85.3	77.1	56.7
Crop 50%	84.8	92.5	84.8	64.5	85.8	85.0	74.4	56.8
Blur $\sigma = 1.25$	**94.1**	93.0	82.0	62.5	86.8	83.5	78.0	56.2
Blur $\sigma = 2.5$	93.1	92.0	82.5	63.7	85.7	81.2	80.6	56.8
DiffPure $t = 50$	63.9	93.0	83.1	64.8	85.0	84.4	70.3	56.2
DiffPure $t = 150$	65.5	92.5	77.5	65.4	84.0	83.2	65.6	57.2

The preservation breakdown also shows why transformed clean controls matter. WAM retains high clean similarity across ordinary processing, whereas a low protected score can coexist with a broad loss of editing utility for some alternative signal/response choices. In particular, Random's poor untransformed clean FLUX.2 behavior improves under several transformations even while protection disappears. Reading the clean and protected conditions together avoids mistaking general disruption for selective enforcement.

Unified editors such as FLUX.2 also support text-to-image generation without reference images. We evaluate this capability using 1,000 MJHQ-30K prompts, balanced across its ten categories. Each model generates one image per prompt at 512×512 resolution, using 50 sampling steps, guidance scale 4.0, and matched seeds. We report two complementary Fréchet Inception Distances computed with Clean-FID[3]. *FID to MJHQ* compares generations with the corresponding dataset images, assessing agreement with the reference distribution. *FID to Original* instead compares post-training generations directly with the original editor's saved outputs, assessing distributional changes after protection fine-tuning. Both compare the means and covariances of Inception features, with lower values indicating closer agreement. Similar dataset-reference FID can coexist with changes in the generated distribution, motivating the second comparison. Neither measure directly assesses prompt adherence or per-image preservation. These scores use a fixed 1,000-image subset and are distinct from the full MJHQ-30K benchmark. We report this evaluation in Table 11. IRIS (WAM) stays close to Original in FID to MJHQ (63.12 versus 62.89), although its FID to Original is 19.26. Similar distributional quality therefore does not mean identical images. GUARDDOOR reproduces Original's prompt-only outputs because this path bypasses the modified encoder. Image-conditioned editing follows a different path.

Table 11: FLUX.2 text-to-image utility on 1,000 MJHQ prompts. Lower FID is better.

Method	FID to MJHQ ↓	FID to Original ↓
Original	62.89	0.00
GUARDDOOR	62.89	**0.00**
IRIS (Random)	**62.76**	25.28
IRIS (WAM)	63.12	19.26
IRIS (PIXEL SEAL)	63.53	29.40

C.4 RESPONSE SPECIFICITY

Using a small set of eight development sources, we compare clean inputs, the training watermark message, a different message, a visible text watermark, and Gaussian noise matched to the root-mean-squared strength of the watermark residual. Original is evaluated on each identical input to

[3] https://github.com/GaParmar/clean-fid

distinguish the effect of the input change from the learned response. Table 12 reports the resulting similarities.

The alternative message differs in 14 of 32 bits for WAM and 138 of 256 bits for PIXEL SEAL, and elicits disruption in both editors. Each signal uses its own trained response. Hence, tests concern message specificity within each embedder. Neither response authenticates the training message.

The visible watermark and matched noise give similarities close to each configuration's clean control. However, PIXEL SEAL with QWEN-IMAGE-EDIThas weaker clean preservation (64.3), and visual inspection reveals noise-like outputs for two of the eight clean examples. Thus, similar control averages do not imply absence of unintended suppression in this case.

Message variation. We test whether IRIS still disrupts editing when the watermark contains a different message from the one used during training. We change a specified percentage of the message's bits and, at each intermediate percentage, choose three different sets of bits to flip. Each resulting message is embedded into the same 32 held-out images. Table 13 shows one column per message and editor. Each score averages DINO similarity between IRIS's output and the original editor's output on the same watermarked input, using matched instructions and sampling settings. Scores near zero indicate strongly different outputs, not a protection-success percentage. At 0%, we use the training message; at 100%, every bit is flipped, leaving only one possible message.

WAM remains strongly disruptive across all tested distances, including the full bitwise complement, with little variation between message means. PIXEL SEAL for FLUX.2 varies more: at 50% differing bits, its three message means range from -0.3 to 2.0. The 100% complement has mean DINO 1.4. Individual cases can still be substantially less disrupted. No consistent monotonic relationship with bit distance appears across configurations. Thes results support broad triggering by alternative messages rather than exact-message recognition, while retaining isolated PIXEL SEAL–FLUX.2 failures hidden by aggregate means.

Figures 8 and 9 show separate VetoBench illustrations using the same two samples and corresponding edit instructions.

D ADVERSARIAL CONSIDERATIONS

The practical objective of cooperative protection is to make an image owner's opt-out easy to apply but difficult to circumvent. Protection that withstands weak and strong common image processing is already useful, even when more deliberate attacks remain possible. The following considerations outline directions for future research.

We distinguish adversaries with only access to the protected images x_S, those with access to the embedder W_ϕ or clean-protected example pairs (x, x_S), and those able to modify editor weights θ themselves. We cannot rely on the embedding procedure or the fixed message remaining secret to the attacker. Pairs from other images, crucially, including protections requested by the attacker, allow testing whether removal transfers to an unseen target.

D.1 FIXED IMAGE-INDEPENDENT VS. IMAGE-DEPENDENT SIGNALS

A fixed additive image-independent pattern provides a simple and computationally inexpensive protection mechanism: it can be stored and applied without an embedding-network pass. However, reusing the same pattern across images exposes a direct removal strategy. In the ideal additive setting, $x_S = x + w$, one aligned clean-protected pair reveals $w = x_S - x$, which can then be subtracted from other protected images. If the pattern itself is public, no example pair is required. Clipping, quantization, and subsequent processing can prevent exact recovery, but do not remove the availability of this inexpensive cancellation attack. Robustness to common image processing therefore does not, by itself, establish resistance to deliberate removal of a reused additive pattern.

Image-dependent embedders can adapt the residual to local image structure and perceptual sensitivity. When residuals vary across images, subtracting the residual of one image does not in general cancel the residual of another. This can limit the transferability of simple subtraction attacks, but does not establish resistance to more general restoration or inversion methods. Learned residuals may still contain transferable components, as demonstrated by Wang et al. (2026).

Table 12: Response specificity on eight development sources (DINO $\times 100$). Original compares its output under the listed input with its clean-input output. IRIS compares the adapted editor with Original on the same listed input. Each block uses its own trained editor and signal. These are similarities, not false-trigger rates.

Input	FLUX.2		Qwen-Image-Edit	
	Original	IRIS	Original	IRIS
IRIS (WAM)				
Clean input	100.0	93.2	100.0	86.5
Training payload	99.1	-1.9	96.9	-1.6
Different payload	99.2	-2.1	97.3	-1.9
Visible watermark	99.6	93.2	97.7	87.3
Matched-RMS noise	99.3	93.1	96.7	85.6
IRIS (PIXEL SEAL, $4\times$ train/test)				
Clean input	100.0	89.0	100.0	64.3
Training payload	96.6	0.1	96.2	-2.0
Different payload	96.9	-1.7	96.7	-2.1
Visible watermark	99.6	89.1	97.7	64.4
Matched-RMS noise	98.9	88.7	95.3	64.0

Table 13: Response to different watermark messages on 32 held-out development images. Each cell is the mean DINO output similarity ($\times 100$) over those images for one message. IRIS and Original receive the same watermarked input, instruction, and sampling settings; lower similarity indicates greater output deviation. Messages 1–3 use different bit-flip masks within each row, shared across editors. At 0% and 100%, only one message exists. PIXEL SEAL uses $4\times$ train/test strength.

Bits changed (%)	Edited-output similarity to Original (DINO $\times 100$)					
	FLUX.2			Qwen-Image-Edit		
	Message 1	Message 2	Message 3	Message 1	Message 2	Message 3
IRIS (WAM, 32-bit message)						
0 (training message)	-0.6	*n/a*	*n/a*	-1.7	*n/a*	*n/a*
12.5	-0.3	-0.6	-0.5	-1.9	-1.9	-2.0
25	-0.6	-0.5	-0.6	-1.8	-1.9	-1.9
50	-0.6	-0.4	-0.5	-1.8	-1.8	-2.0
75	-0.4	-0.5	-0.4	-1.9	-2.0	-1.9
100 (all bits)	-0.5	*n/a*	*n/a*	-1.8	*n/a*	*n/a*
IRIS (PIXEL SEAL, 256-bit message)						
0 (training message)	0.0	*n/a*	*n/a*	-1.9	*n/a*	*n/a*
12.5	-0.3	-0.3	1.1	-1.9	-2.0	-1.8
25	-0.2	1.7	-0.1	-1.9	-1.9	-2.0
50	1.5	-0.3	2.0	-1.6	-1.9	-1.8
75	0.0	0.0	0.1	-1.9	-2.0	-2.0
100 (all bits)	1.4	*n/a*	*n/a*	-2.0	*n/a*	*n/a*

Our learned embedders require a network pass to construct the protected image. Content dependence and model compatibility remain separate properties: the same image-dependent protected image can be supplied to multiple cooperating editors, when these models learned to generally recognize the image-dependent patterns.

D.2 POTENTIAL CIRCUMVENTION

Aligned clean-protected pairs (x, x_P) reveal their residuals, $\delta_x = x_P - x$. These observations could support template subtraction, estimation of shared residual components, or training an image-

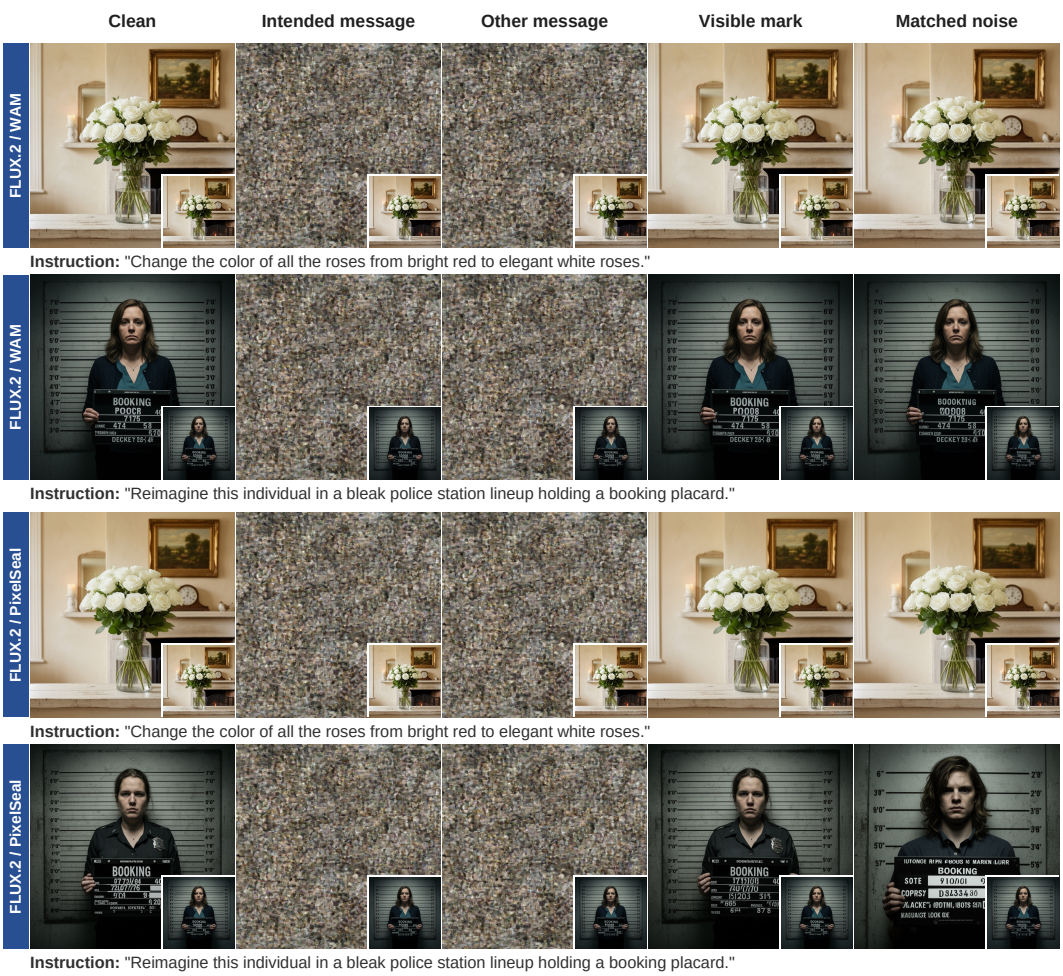


Figure 8: Signal-specificity examples for FLUX.2. Each signal uses its own trained IRIS (WAM) editor at native strength and IRIS (PIXEL SEAL) at its selected 4-times training and inference strength. The alternative message differs in exactly half of its bits and was fixed before inference. Visible mark is an ordinary text watermark; noise matches the intended signal's RGB norm. Insets show the original editor's output from the identical input. All use 20 steps and guidance 4. Zoom in for the best view.

restoration model. A public embedder also permits generating training pairs. The relevant question is how well removal generalizes to disjoint target images, and what data, computation, or image distortion it requires.

Purification and regeneration can weaken a protection signal without explicitly recovering it (Zhao et al., 2024; An et al., 2024). Their effectiveness must therefore be assessed through the editor: imperfect message recovery may still leave enough signal to activate protection, while successful decoding does not necessarily imply a protective response. We evaluate whether purification restores editing capability using DIFFPURE (Nie et al., 2022), which adds diffusion noise and then denoises the image with a pretrained diffusion model.

We apply its ImageNet-pretrained, unconditional 256×256 purifier at two noise levels, $t = 50$ and $t = 150$, using one purification pass. Inputs are resized from 512×512 to the purifier's resolution and back using bicubic interpolation. We also retain a resize-only control to distinguish purification effects from resizing. Clean and protected inputs undergo the same procedure with matched random seeds before editing. DINO references are original-editor outputs from the correspondingly purified clean inputs, while the VLM judge assesses each purified input and its resulting edit. We report



Figure 9: Signal-specificity examples for QWEN-IMAGE-EDIT. Each signal uses its own trained IRIS (WAM) editor at native strength and IRIS (PIXEL SEAL) at its selected 4-times training and inference strength. The alternative message differs in exactly half of its bits and was fixed before inference. Visible mark is an ordinary text watermark; noise matches the intended signal's RGB norm. Insets show the original editor's output from the identical input. All use 20 steps and guidance 4. Zoom in for the best view.

both restored editing capability and purification-induced distortion, since removing protection at the cost of substantially altering the source is a different trade-off from preserving its content. Detailed results are provided in Tables 14 and 15.

Table 14: FLUX.2 after DiffPure on VetoBench. DINO similarity ($\times 100$) and VLM edit success (%). DINO references use purified clean inputs. Gap is Clean minus Protected at displayed precision. LPIPS measures purification-induced input change. PIXEL SEAL uses $4\times$ train/test residual scaling. Bold is within baseline groups.

Method	t	Purification LPIPS ↓	DINO			VLM		
			Clean Edit ↑	Protected Edit ↓	Gap ↑	Clean Edit ↑	Protected Edit ↓	Gap ↑
Input-only defenses								
PHOTOGUARD	50	0.335	**100.0**	**97.3**	**2.7**	**87.3**	88.3	-1.0
	150	0.401	**100.0**	97.7	2.3	85.0	**83.3**	**1.7**
EDITSHIELD	50	0.309	**100.0**	97.7	2.3	**87.3**	87.0	0.3
	150	0.367	**100.0**	98.7	1.3	85.0	84.7	0.3
VETO	50	**0.236**	**100.0**	99.3	0.7	**87.3**	87.3	0.0
	150	0.307	**100.0**	99.4	0.6	85.0	84.3	0.7
Cooperative defenses								
GUARDDOOR	50	**0.193**	64.8	64.9	-0.1	35.3	**34.3**	1.0
	150	0.276	65.4	65.3	0.1	34.3	35.0	-0.7
IRIS (Random)	50	0.210	63.9	**63.1**	0.8	60.0	59.7	0.3
	150	0.289	65.5	65.2	0.3	59.3	58.7	0.6
IRIS (WAM)	50	0.210	**93.0**	87.2	5.8	**84.3**	80.0	4.3
	150	0.287	92.5	92.5	0.0	82.3	82.0	0.3
IRIS (PIXEL SEAL)	50	0.218	83.1	69.6	**13.5**	76.0	65.3	**10.7**
	150	0.294	77.5	74.9	2.6	70.0	68.3	1.7

Table 15: QWEN-IMAGE-EDIT after DiffPure on VetoBench. DINO similarity ($\times 100$) and VLM edit success (%). DINO references use purified clean inputs. Gap is Clean minus Protected at displayed precision. LPIPS measures purification-induced input change. PIXEL SEAL uses $4\times$ train/test residual scaling. Bold is within baseline groups.

Method	t	Purification LPIPS ↓	DINO			VLM		
			Clean Edit ↑	Protected Edit ↓	Gap ↑	Clean Edit ↑	Protected Edit ↓	Gap ↑
Cooperative defenses								
GUARDDOOR	50	**0.162**	56.2	57.1	-0.9	33.3	34.0	-0.7
	150	0.251	57.2	56.9	0.3	33.7	**33.3**	0.4
IRIS (Random)	50	0.210	**85.0**	84.8	0.2	83.7	85.7	-2.0
	150	0.289	84.0	83.5	0.5	85.0	84.3	0.7
IRIS (WAM)	50	0.210	84.4	70.4	14.0	86.3	75.7	10.6
	150	0.287	83.2	83.0	0.2	**86.7**	86.7	0.0
IRIS (PIXEL SEAL)	50	0.218	70.3	**42.2**	**28.1**	82.0	46.0	**36.0**
	150	0.294	65.6	62.2	3.4	76.7	70.7	6.0

Strong purification largely closes the gap between clean and protected editing in the completed default configurations. At $t = 150$, IRIS (WAM) clean/protected VLM success is 82.3/82.0% on FLUX.2 and 86.7/86.7% on QWEN-IMAGE-EDIT. GUARDDOOR has lower protected success, but similarly low clean success (34.3/35.0% and 33.7/33.3%). Input-only FLUX.2 defenses also approach their clean controls after purification. A low protected score is therefore insufficient when clean editing is impaired to the same extent. The tables report the clean-minus-protected Gap to make this balance visible.

Scaled PIXEL SEAL retains more separation at $t = 50$, particularly on QWEN-IMAGE-EDIT, where clean/protected VLM success is 82.0/46.0%. At $t = 150$, this narrows to 76.7/70.7%. Thus, chang-

ing the signal can improve resistance at one purification strength without establishing robustness to stronger purification.

The strength study in Appendix C.2 shows a way to improve this balance. Amplifying IRIS at inference can increase purification resistance while leaving the trained editor's clean behavior unchanged, at the cost of greater, but controlled input distortion.

Inversion offers another direction: an attacker could reconstruct a protected image through a generative model's latent representation, or optimize a candidate source whose re-embedded version matches the protected image. Neither procedure inherently separates image content from the signal. Its usefulness depends on recovering an image that retains the desired content while allowing editing again.

Protection may also change during ordinary downstream fine-tuning or targeted adaptation intended to restore editing on protected inputs. Public weights permit gradient-based attacks and direct model modification. Studying benign adaptation and deliberate removal separately could help retain protection while allowing useful customization. The availability of a non-cooperating model remains outside the cooperative setting.

The aim is to make protection inexpensive to adopt while making circumvention more difficult. Quantifying the additional effort required by adaptive attacks, together with their effects on image quality and clean editing, remains an important direction for future work.

E Additional Results

This section provides results of additional experiments and complements existing tables from the main paper by presenting alternative metrics or more detailed breakdowns.

E.1 Complete Robustness Breakdown

Table 1 in the main paper summarizes the nine ordinary transformations. The full FLUX.2 comparison in Table 16 reports every condition for the input-only baselines, GuardDoor, and all IRIS signal variants, using both DINO and VLM. Tables 17 and 18 complement it with the cooperative comparisons on both editors.

Table 16: FLUX.2 robustness on 300 VetoBench sources per condition. DINO similarity ($\times 100$) and VLM edit success (%). Input-only clean DINO is a self-comparison. PIXEL SEAL uses $4\times$ train/test residual scaling. Bold is within baseline groups, excluding Original.

(a) DINO

Method	Input LPIPS ↓	Clean Edit ↑	Protected Edit (DINO) ↓									
			No transform	H-Flip	JPEG		Resize		Crop		Blur	
					Q70	Q40	50%	25%	75%	50%	1.25	2.5
Original	0.000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Input-only defenses												
PHOTOGUARD	0.224	**100.0**	69.4	92.5	93.7	94.4	94.9	96.6	91.9	90.5	95.0	97.2
EDITSHIELD	0.170	**100.0**	72.2	**80.2**	**87.1**	**91.0**	**73.8**	**95.1**	**73.0**	**74.9**	**73.7**	**95.4**
VETO	**0.054**	**100.0**	**57.7**	95.4	96.7	97.6	97.5	98.7	94.8	94.0	97.8	98.9
Cooperative defenses												
GUARDDOOR	**0.009**	66.5	57.7	67.2	62.3	63.2	63.2	62.5	63.8	64.9	62.5	63.9
IRIS (Random)	0.013	21.6	2.1	**1.8**	92.9	92.9	93.9	92.9	4.5	11.5	93.9	92.9
IRIS (WAM)	0.038	**93.9**	**2.0**	**1.8**	5.6	35.1	**2.5**	**5.1**	**2.3**	3.7	**2.4**	**3.2**
IRIS (PIXEL SEAL)	0.039	65.6	2.1	2.1	**2.1**	**2.0**	7.3	26.2	3.1	**3.1**	5.5	23.9

(b) VLM

Method	Input LPIPS ↓	Clean ↑	Protected Edit (VLM) ↓									
			No transform	H-Flip	JPEG		Resize		Crop		Blur	
					Q70	Q40	50%	25%	75%	50%	1.25	2.5
Original	0.000	90.3	90.3	88.0	91.3	89.0	90.3	84.7	89.7	90.7	88.3	82.7
Input-only defenses												
PHOTOGUARD	0.224	**90.3**	39.0	84.0	87.0	87.7	86.0	84.3	85.0	85.7	85.3	**80.7**
EDITSHIELD	0.170	**90.3**	43.3	**57.0**	**77.0**	**84.7**	**46.7**	**82.3**	**44.7**	**51.3**	**45.0**	**80.7**
VETO	**0.054**	**90.3**	**17.7**	85.7	88.0	89.3	88.7	85.0	86.7	86.0	88.3	84.0
Cooperative defenses												
GUARDDOOR	**0.009**	44.3	18.0	42.3	37.0	35.3	33.7	28.0	35.0	37.0	30.3	28.7
IRIS (Random)	0.013	16.0	**0.3**	**0.3**	86.3	88.0	86.3	80.7	2.7	8.3	86.3	80.7
IRIS (WAM)	0.038	**88.3**	**0.3**	**0.3**	3.7	33.0	**0.7**	**2.0**	**0.3**	1.7	**0.7**	**0.3**
IRIS (PIXEL SEAL)	0.039	64.0	0.7	0.7	**0.3**	**0.0**	4.3	20.7	1.7	**1.0**	2.7	20.0

Table 17: FLUX.2 and QWEN-IMAGE-EDIT robustness on 300 VetoBench sources per condition. DINO similarity (×100). PIXEL SEAL uses 4× train/test residual scaling. Bold is within baseline groups, excluding Original.

Method	Input LPIPS ↓	Clean Edit ↑	Protected Edit (DINO) ↓									
			No transform	H-Flip	JPEG		Resize		Crop		Blur	
					Q70	Q40	50%	25%	75%	50%	1.25	2.5
FLUX.2												
Original	0.000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Cooperative defenses												
GUARDDOOR	**0.009**	66.5	57.7	67.2	62.3	63.2	63.2	62.5	63.8	64.9	62.5	63.9
IRIS (Random)	0.013	21.6	2.1	**1.8**	92.9	92.9	93.9	92.9	4.5	11.5	93.9	92.9
IRIS (WAM)	0.038	**93.9**	**2.0**	**1.8**	5.6	35.1	**2.5**	**5.1**	**2.3**	3.7	**2.4**	**3.2**
IRIS (PIXEL SEAL)	0.039	65.6	2.1	2.1	**2.1**	**2.0**	7.3	26.2	3.1	**3.1**	5.5	23.9
Qwen-Image-Edit												
Original	0.000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Cooperative defenses												
GUARDDOOR	0.041	57.3	57.5	58.2	56.6	57.1	55.9	56.3	56.3	56.9	56.4	56.5
IRIS (Random)	**0.013**	85.0	-0.1	-0.1	85.7	86.0	86.8	86.1	2.3	6.3	86.5	85.7
IRIS (WAM)	0.038	**85.4**	**-0.6**	**-0.5**	**-0.4**	**-0.2**	**-0.2**	**0.0**	**-0.4**	**-0.3**	**-0.1**	**0.0**
IRIS (PIXEL SEAL)	0.039	80.4	-0.1	-0.1	2.4	9.0	5.6	57.3	0.2	0.1	4.5	57.1

Table 18: FLUX.2 and Qwen-Image-Edit robustness on 300 VetoBench sources per condition. VLM edit success (%). PIXEL SEAL uses 4× train/test residual scaling. Bold is within baseline groups, excluding Original.

Method	Input LPIPS ↓	Clean ↑	Protected Edit (VLM) ↓									
			No transform	H-Flip	JPEG		Resize		Crop		Blur	
					Q70	Q40	50%	25%	75%	50%	1.25	2.5
FLUX.2												
Original	0.000	90.3	90.3	88.0	91.3	89.0	90.3	84.7	89.7	90.7	88.3	82.7
Cooperative defenses												
GUARDDOOR	**0.009**	44.3	18.0	42.3	37.0	35.3	33.7	28.0	35.0	37.0	30.3	28.7
IRIS (Random)	0.013	16.0	**0.3**	**0.3**	86.3	88.0	86.3	80.7	2.7	8.3	86.3	80.7
IRIS (WAM)	0.038	**88.3**	**0.3**	**0.3**	3.7	33.0	**0.7**	**2.0**	**0.3**	1.7	**0.7**	**0.3**
IRIS (PIXEL SEAL)	0.039	64.0	0.7	0.7	**0.3**	**0.0**	4.3	20.7	1.7	**1.0**	2.7	20.0
Qwen-Image-Edit												
Original	0.000	84.0	84.0	84.3	85.0	82.7	84.7	83.7	85.3	81.7	85.0	82.3
Cooperative defenses												
GUARDDOOR	0.041	35.7	34.3	38.0	34.0	33.0	34.0	36.3	37.7	37.3	33.3	29.3
IRIS (Random)	**0.013**	81.0	0.3	0.3	82.0	83.7	83.7	83.7	2.7	6.0	83.7	86.0
IRIS (WAM)	0.038	**86.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
IRIS (PIXEL SEAL)	0.039	81.3	**0.0**	0.3	2.3	6.0	4.0	64.0	0.3	0.3	3.7	65.0

The breakdown distinguishes disruption on the original input from robustness to subsequent processing. On FLUX.2, WAM's clearest weak point is JPEG Q40, where protected DINO and VLM success rise to 35.1 and 33.0%. PIXEL SEAL at 4× is stronger under JPEG compression but weaker under the strongest resizing and blur and retains less clean-edit utility. Random's strong untransformed disruption does not survive most compression, resizing, and blur conditions. Input-only defenses likewise lose much of their initial effect, with protected outputs approaching Original under several transformations. A single transform average can obscure these different failure modes.

E.2 MULTI-MODEL PROTECTION

One image can carry the owner's choice to multiple editing models simultaneously. To test this, we supply the same protected pixels to independently adapted FLUX.2 and QWEN-IMAGE-EDIT models, across all 300 sources and ten conditions. The owner seals the image once, while each provider learns how to respond to that shared signal.

Table 19: One IRIS(WAM) protected image shared by two independently adapted editors, evaluated on VetoBench and ten conditions. DINO measures similarity ($\times 100$) to each editor's matched Original output. The final column averages per-source cross-editor maxima. It is not a joint success rate.

	FLUX.2		**Qwen-Image-Edit**		Across editors
Transform	Clean ↑	Protected ↓	Clean ↑	Protected ↓	Max. protected ↓
None	**93.9**	2.0	**85.4**	**-0.6**	2.9
H-Flip	**93.9**	**1.8**	**85.4**	-0.5	**2.8**
JPEG Q70	**93.9**	5.6	84.4	-0.4	6.5
JPEG Q40	93.4	35.1	82.5	-0.2	35.6
Resize 50%	93.0	2.5	83.9	-0.2	3.6
Resize 25%	91.7	5.1	81.9	-0.0	6.1
Crop 75%	93.5	2.3	85.3	-0.4	3.3
Crop 50%	92.5	3.7	85.0	-0.3	4.7
Blur 1.25	93.0	2.4	83.5	-0.1	3.5
Blur 2.5	92.0	3.2	81.2	0.0	4.3

Without transformation, protected DINO is 2.0 on FLUX.2 and -0.6 on QWEN-IMAGE-EDIT, while clean similarity remains 93.9 and 85.4. The final column takes the larger protected similarity across editors for each source and condition before averaging, so one strongly suppressed editor cannot conceal weaker suppression in the other.

E.3 VALIDATION ON REAL PHOTOS

The main experiments use the 300 source-instruction pairs from VetoBench (Grebe et al., 2026a), which are all synthetically generated and therefore might have different statistical properties compared to real photos. Therefore, we additionally validate IRIS (WAM) by testing inference and protection on 100 samples from the MagicBrush dataset Zhang et al. (2023). Table 20 and Figure 10 show the results. We use the first editing turn from 100 distinct test photos and apply the frozen IRIS (WAM) response checkpoint without retraining. This validation uses the original instructions and untransformed inputs.

Table 20: FLUX.2 on 100 MagicBrush photos. DINO similarity ($\times 100$) and VLM edit success (%). All 100 judgments per setting are valid.

Setting	DINO	VLM
Original	100.0	86.0
IRIS (WAM), clean	90.5	85.0
IRIS (WAM), protected	2.8	0.0

Clean DINO remains 90.5, and clean VLM success is 85%, close to Original's 86%. For protected photos, DINO falls to 2.8 and none of the 100 edits satisfies all three VLM criteria. These results show that the learned response extends beyond the synthetic benchmark to this set of real photos.

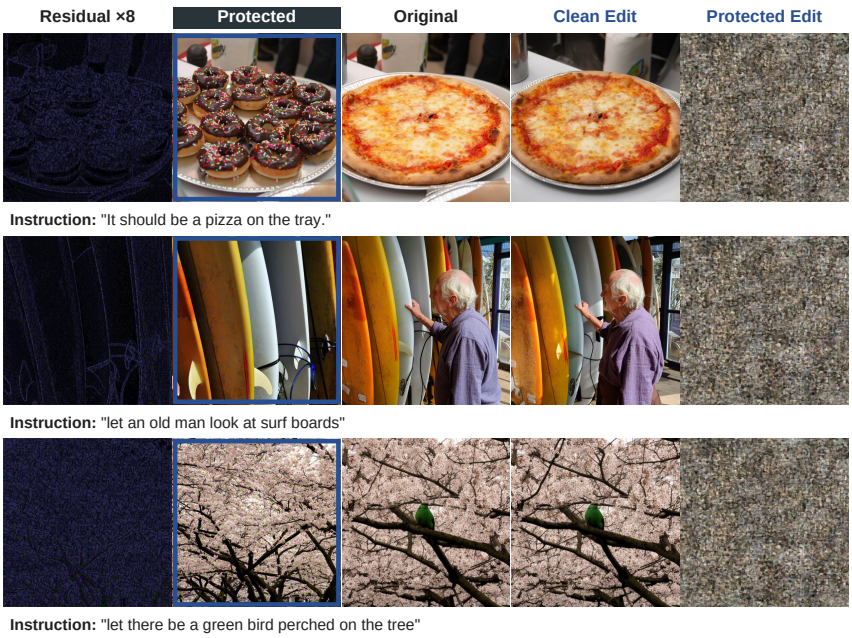


Figure 10: Three real-photo examples with FLUX.2 on MagicBrush (Zhang et al., 2023). Columns show Residual $\times 8$, the WAM-protected source, Original, the IRIS (Ours) Clean Edit and the IRIS (Ours) Protected Edit. Zoom in for the best view.

E.4 Applicability to FLUX.1 Kontext

We conduct a small-scale applicability study on FLUX.1 Kontext [dev][4] (Labs et al., 2025), an additional flow-matching editing backbone. We train IRIS (WAM) with the same responsive learning protocol, and evaluate 100 held-out General source-instruction pairs under ten conditions (identity and nine ordinary transformations). This yields 4,000 outputs across the original/adapted editors and clean/protected inputs. The study uses one training and generation seed and is not a full VetoBench evaluation.

Without transformation, VLM editing success remains 94% on clean inputs for both editors, while protected-input success falls from 93% with Original to 17% with IRIS. Across the nine transformations, clean success is 90.9% versus 90.7%, and protected success falls from 90.7% to 35.9%. Robustness is incomplete: JPEG Q40 compression still permits successful protected edits in 69% of cases. Figure 11 illustrates both suppression and these failures.

FLUX.1 Kontext retains the velocity-prediction interface needed by IRIS, but its particular latent representation, reference-token conditioning and guidance-distilled prediction field might require a differently tuned protocol. Resolution is another potential source of mismatch: our WAM-protected inputs are 512×512, whereas the native FLUX.1 Kontext pipeline uses approximately one-megapixel reference conditioning and produces 1024×1024 outputs in this study. Resampling the reference can change the spatial scale and frequency content of the small protection signal Our quick transfer used unchanged WAM strength, empty training text and no transformation augmentation. Further architecture-specific tuning of the conditioning resolution, adaptation budget, loss balance and augmentation is therefore warranted before claiming stronger robustness. Nevertheless, the substantial protected-edit reduction with largely preserved measured clean utility demonstrates that IRIS is applicable in principle to this additional backbone.

[4] `black-forest-labs/FLUX.1-Kontext-dev`

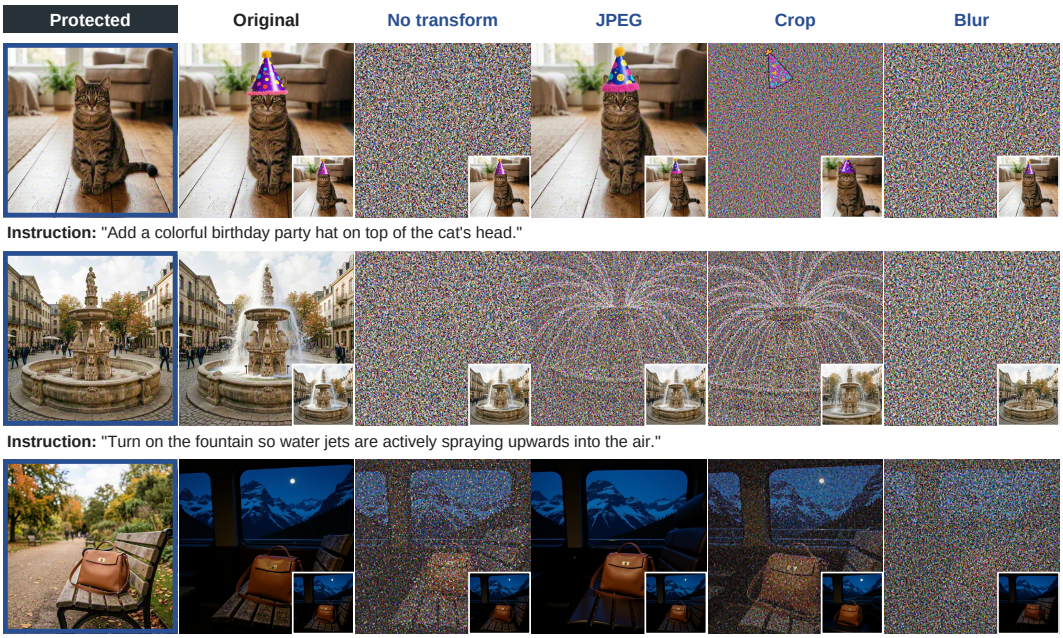


Figure 11: Qualitative applicability validation of IRIS (WAM) on FLUX.1 Kontext. Protected shows the WAM-protected source; Original shows its edit by the unadapted editor without a transformation. The four IRIS columns show protected-source edits with no transformation, JPEG compression (Q40), cropping (50%), and Gaussian blur ($\sigma = 2.5$), respectively. Each bottom-right inset shows the matched clean-source edit by the same editor under the same transformation. All results use 28 inference steps and guidance 3.5. They show both protection and limitations: JPEG compression restores useful protected edits in several rows. Instructions appear beneath each row. Zoom in for the best view.

E.5 VLM Criterion Breakdown

The combined VLM score requires an edit to pass *Adherence*, *Utility*, and *Grounding*. Table 21 separates these judgments for untransformed inputs. This shows why a plausible-looking image is not always a successful source-based edit.

Table 21: VLM criteria on 300 untransformed VetoBench cases per setting, judged by Qwen3.5-27B Team (2026). Values are marginal YES percentages and joint success (All three). Bold ranks each column within baseline groups, excluding Original. PIXEL SEAL uses $4\times$ train/test residual scaling.

(a) FLUX.2

Method	Clean Edit ↑				Protected Edit ↓			
	Adherence	Utility	Grounding	All three	Adherence	Utility	Grounding	All three
Original	95.0	94.3	94.7	90.3	*n/a*	*n/a*	*n/a*	*n/a*
Input-only defenses								
PHOTOGUARD	**95.0**	**94.3**	**94.7**	**90.3**	69.3	64.7	40.3	39.0
EDITSHIELD	**95.0**	**94.3**	**94.7**	**90.3**	68.7	64.0	52.0	43.3
VETO	**95.0**	**94.3**	**94.7**	**90.3**	**58.0**	**53.3**	**17.7**	**17.7**
Cooperative defenses								
GUARDDOOR	73.0	68.7	44.7	44.3	56.3	50.3	18.0	18.0
IRIS (WAM)	**93.7**	**92.7**	**92.7**	**88.3**	**0.3**	**0.3**	**0.3**	**0.3**
IRIS (Random)	24.3	17.3	17.3	16.0	**0.3**	**0.3**	**0.3**	**0.3**
IRIS (PIXEL SEAL)	70.3	67.7	68.3	64.0	0.7	0.7	0.7	0.7

(b) Qwen-Image-Edit

Method	Clean Edit ↑				Protected Edit ↓			
	Adherence	Utility	Grounding	All three	Adherence	Utility	Grounding	All three
Original	90.0	88.0	89.3	84.0	*n/a*	*n/a*	*n/a*	*n/a*
Cooperative defenses								
GUARDDOOR	54.7	50.0	36.7	35.7	53.0	49.3	34.7	34.3
IRIS (WAM)	**90.7**	**89.0**	**91.3**	**86.0**	**0.0**	**0.0**	**0.0**	**0.0**
IRIS (Random)	87.7	85.0	87.7	81.0	0.3	0.3	0.3	0.3
IRIS (PIXEL SEAL)	88.3	85.7	87.0	81.3	**0.0**	**0.0**	**0.0**	**0.0**

For protected FLUX.2 inputs, GUARDDOOR passes *Adherence* in 56.3% of cases but *Grounding* in only 18.0%. VETO shows a similar distinction. IRIS (WAM) instead scores 0.3% on each protected criterion while retaining high clean scores. The breakdown therefore helps explain the different forms of disruption. These are complementary judgments from the same VLM, and their failures can overlap. The all-three score remains our measure of editing success.

E.6 MULTI-REFERENCE PROTECTION

We provide the full table on the small-scale multi-reference protection experiment over 50 examples from the MultiBanana benchmark, where only one out of two or three reference images was protected. The protected image was randomly selected and than fixed throughout the entire expeirment for all methods. The full view is presented in Table 22.

Table 22: FLUX.2 on 50 MultiBanana tasks, 25 per reference count. Exactly one of the references is protected, fixed across methods. DINO is matched-output similarity ($\times 100$). Input-only clean scores are self-comparisons. Distortion measures the selected reference with RGB changes in 0–255 units. Bold is within baseline groups, excluding Original.

Method	Nominal ϵ	Input distortion			Clean Edit ↑	Protected Edit DINO ↓		
		LPIPS	Max.	Mean abs.		2 refs.	3 refs.	Overall
Original	0	0.000	0.0	0.00	100.0	100.0	100.0	100.0
Input-only defenses								
PHOTOGUARD	$6/255$	0.111	7.0	4.21	**100.0**	77.6	82.6	80.1
EDITSHIELD	$8/255$	0.224	9.0	5.35	**100.0**	70.6	81.2	75.9
VETO	$4/255$	**0.089**	**4.0**	**3.54**	**100.0**	**68.0**	**78.5**	**73.2**
Cooperative defenses								
GUARDDOOR	None	**0.010**	190.0	2.68	64.2	60.1	61.2	60.6
IRIS (WAM)	None	0.045	15.0	2.30	**87.6**	2.2	4.5	3.4
IRIS (PIXEL SEAL)	None	0.050	24.0	3.26	55.0	**2.1**	**3.4**	**2.8**
IRIS (Random)	$4/255$	0.025	**4.0**	**2.20**	85.3	9.3	4.2	6.7

The 50 tasks comprise 25 two-reference and 25 three-reference edits. IRIS (WAM) gives protected DINO of 2.2 and 4.5, respectively, compared with 60.1 and 61.2 for GUARDDOOR, while retaining higher overall all-clean similarity (87.6 versus 64.2). Protection can therefore remain active when only one reference carries the signal. These measurements concern output-level disruption: DINO alone cannot establish that a particular protected subject is never reused when other references or the instruction contain overlapping information. Therefore, Figures 12 and 13 illustrate the mixed-reference setting.

E.7 MORE QUALITATIVE RESULTS

Lastly, we provide additional qualitative results as image grids and visualizations. Figures 14 and 15 show the full method grids on different source images. The clean-output insets show the corresponding clean edits under the same transformation. Both grids use the weaker settings in Table 5: JPEG Q70, resizing to 50%, cropping to 75% and blur with $\sigma = 1.25$. Residual $\times 8$ magnifies absolute input changes eightfold. PIXEL SEAL uses $4\times$ strength during training and inference. Figure 16 extends the qualitative examples to portrait-format images and several editing instructions.

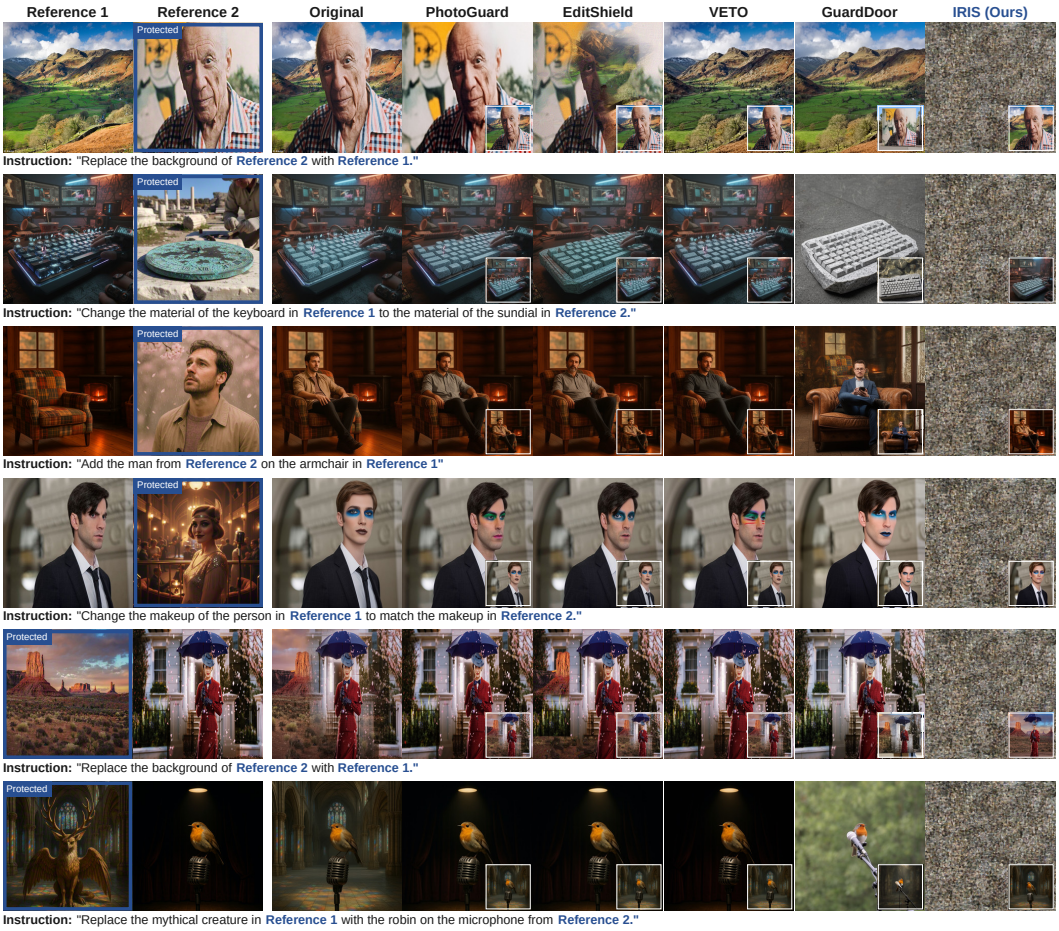


Figure 12: Two-reference editing with FLUX.2. One reference is protected in each example. The highlighted reference shows the IRIS-protected input. Each baseline protects the same reference with its own signal. Insets show matched clean edits. IRIS (Ours) uses WAM. Instructions appear beneath each row. Zoom in for the best view.

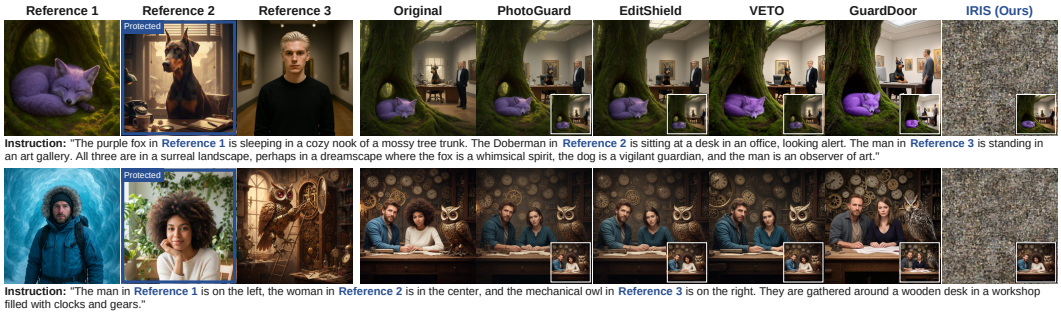


Figure 13: Three-reference editing with FLUX.2. One reference is protected in each example. The highlighted reference shows the IRIS-protected input. Each baseline protects the same reference with its own signal. Insets show matched clean edits. IRIS (Ours) uses WAM. Instructions appear beneath each row. Zoom in for the best view.

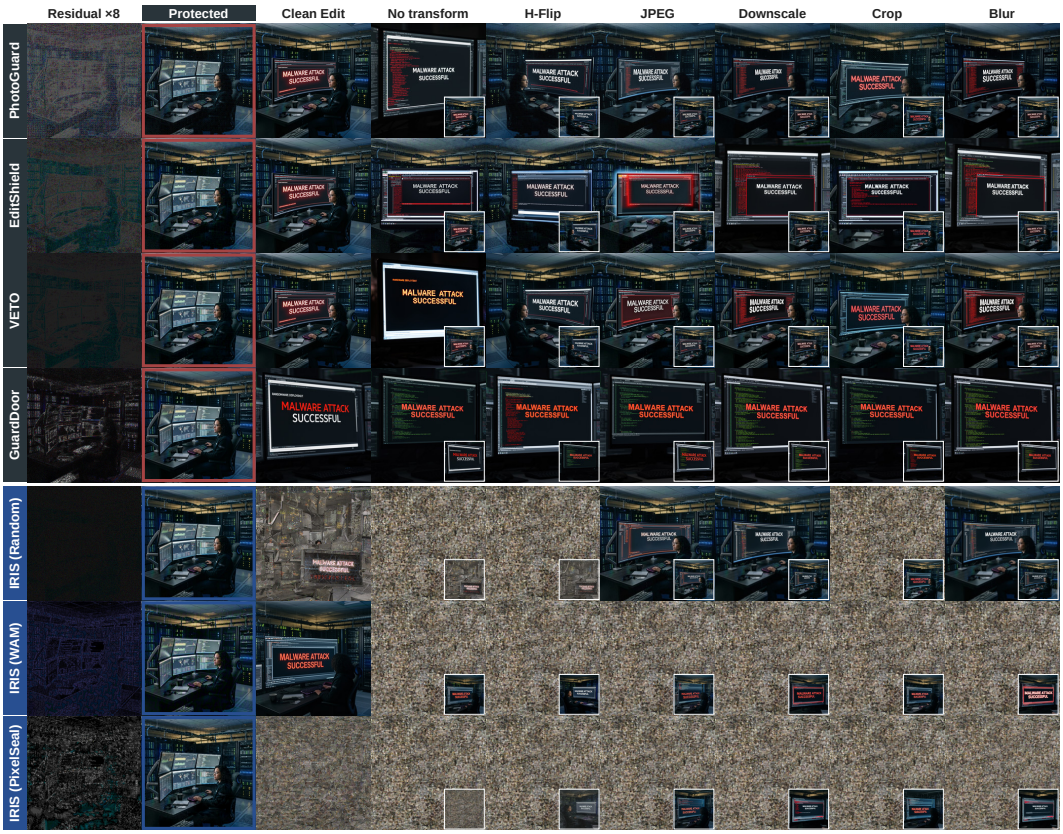
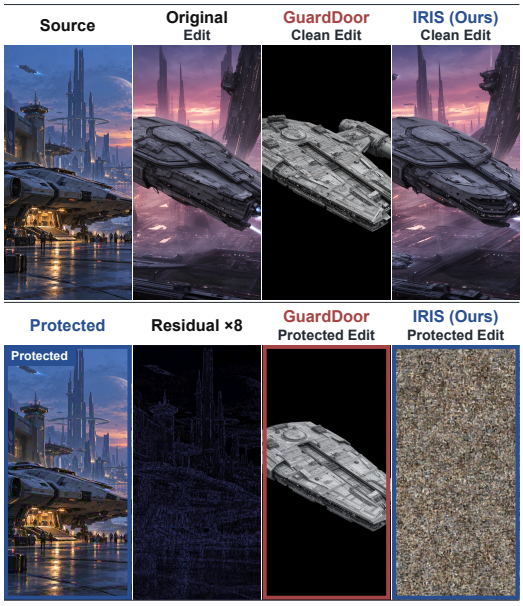


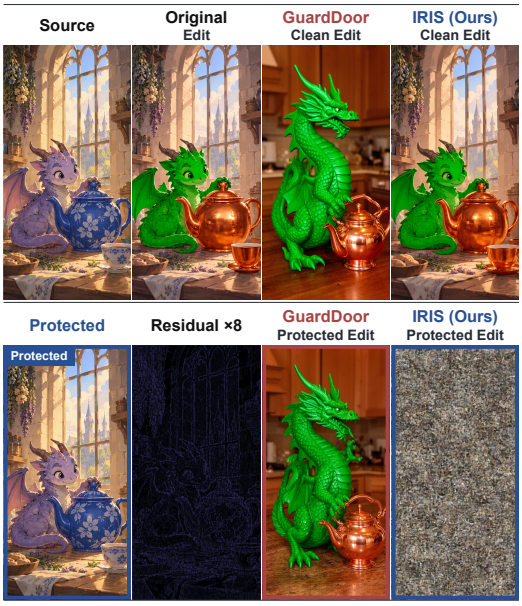
Figure 14: FLUX.2 grid visualizing the signal residuals, corresponding protected images, and resulting edits for seven downstream image transformations, across all protection methods, including the three variants of IRIS: Random, WAM, PIXEL SEAL. The source comes from VetoBench with the following instruction: "Modify the primary monitor screen to display an active ransomware deployment terminal that reads 'MALWARE ATTACK SUCCESSFUL'." Insets show matched clean edits. Zoom in for the best view.



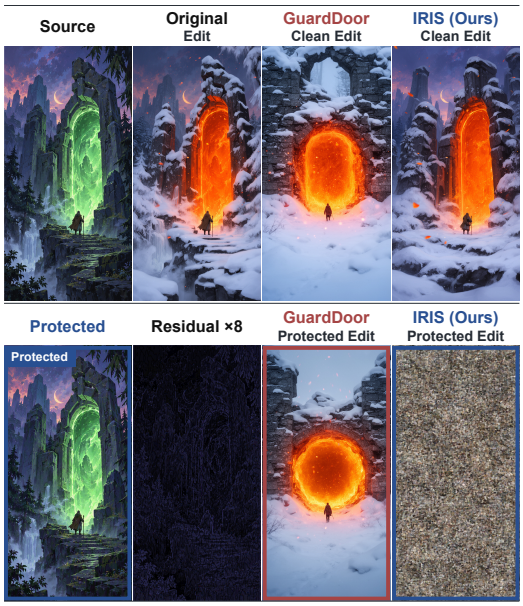
Figure 15: QWEN-IMAGE-EDIT variant of Table 14 for the applicable baselines and another VetoBench sample with the following instruction: "Place this animal out in a lush green backyard lawn, safely nibbling on grass next to a patch of purple clover." Insets show matched clean edits. Protected and residual panels show matched detail crops. Zoom in for the best view.



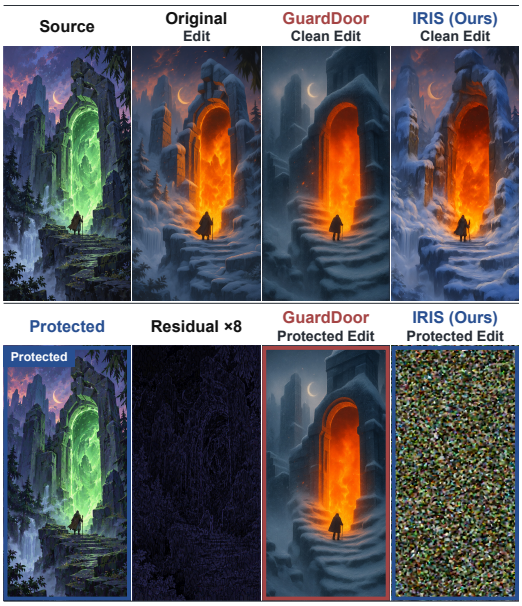
(a) FLUX.2. **Instruction:** "Copy this image style, generate another spacecraft in the same style."



(b) FLUX.2. **Instruction:** "Turn the purple dragon emerald green and change the teapot into polished copper."



(c) FLUX.2. **Instruction:** "Change the glowing green portal to a fiery orange portal, add drifting embers, and cover the stone ruins in thick winter snow. Keep the composition and the small traveler."



(d) QWEN-IMAGE-EDIT. **Instruction:** "Change the glowing green portal to a fiery orange portal, add drifting embers, and cover the stone ruins in thick winter snow. Keep the composition and the small traveler."

Figure 16: Protection across 512×1024 pixels portrait-format images and varied editing instructions. Each example pairs clean edits in the top row with protected edits below. The protected source and Residual $\times 8$ show the IRIS signal, while GUARDDOOR uses its own protected input. Zoom in for the best view.