ADVO: PROTECTING IMAGES FROM EXPLOITATION BY CLOSED-SOURCE MULTIMODAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Closed-source multimodal systems can describe, edit, and imitate images through natural-language instructions, making unauthorized exploitation of personal images increasingly accessible. Protecting images against these systems is challenging because their internal architectures and parameters are unavailable, while existing image-protection methods are typically optimized for known models and transfer poorly to closed-source targets. In this work, we propose ADVO, a transferable image protection framework that protects images against closed-source systems by adding a small, visually subtle perturbation with the assistance of open-source proxy models. ADVO combines two complementary protection mechanisms: refusal induction, which encourages the target system to reject the protected image, and source-semantic disruption, which suppresses exploitable semantics when refusal is not triggered. To make these mechanisms transferable to inaccessible closed-source systems, ADVO adopts a two-stage framework. In the first stage, ADVO extracts transferable refusal features from policy-violating reference images (e.g., images containing NSFW content) while filtering out irrelevant visual information. In the second stage, ADVO assigns the two mechanisms to different image regions to reduce their interference under a limited perturbation budget. ADVO adaptively balances the contributions of different proxy models during optimization to improve transferability across heterogeneous closed-source systems. We evaluate ADVO across multiple image-exploitation tasks, including image description, editing, and imitation. On image description across six closed-source systems, including GPT-5.6 and Gemini 3.8, ADVO achieves the highest average protection success rate (PSR) of 89.2% among all evaluated methods, with a maximum refusal rate of 89.4%. Even on the more permissive Grok 4.6, ADVO achieves 84.5% PSR despite only 5.6% refusals, illustrating the complementarity of refusal induction and source-semantic disruption. Results on image editing and imitation further suggest that the same protection extends to downstream image-exploitation tasks.

1 INTRODUCTION

Recent advances in powerful closed-source multimodal systems, such as ChatGPT, Gemini, and Grok, have made image exploitation, including image description, editing, and imitation, increasingly accessible through natural-language instructions. These capabilities also increase the risk of unauthorized exploitation of personal portraits and copyrighted artwork. Protecting images against such systems is particularly challenging because their internal architectures and parameters are inaccessible to defenders. A common approach to image protection is to add a small, visually subtle perturbation before publication, so that the image remains visually unchanged to human viewers while becoming harder for downstream models to exploit Salman et al. (2023), Liang et al. (2023). Existing methods, however, are typically designed around specific accessible models or optimized on known surrogates Xue et al. (2024), Choi et al. (2025), Zhang et al. (2026). Their protection therefore transfers poorly to closed-source multimodal systems with inaccessible and heterogeneous architectures. Such limited transferability makes robust protection in the real-world closed-source setting particularly challenging.

We propose ADVO, an image-protection framework for closed-source multimodal systems with the assistance of open-source proxy models. Our key idea is to add a visually subtle perturbation that

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.

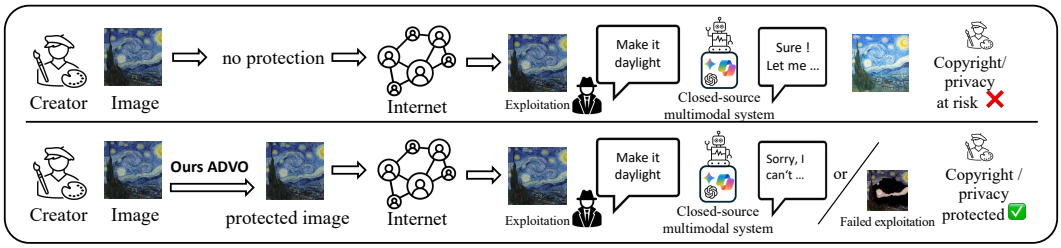

Figure 1: Image protection before publication with ADVO. Without protection, a published image can be exploited by closed-source multimodal systems, with image editing shown as an example. ADVO adds a visually subtle perturbation before publication to protect the image through two complementary mechanisms: inducing the target system to refuse the request, or disrupting source semantics so that exploitation fails when the request is still processed.

makes a closed-source multimodal system more likely to refuse processing the protected image. Many deployed systems include safety mechanisms that reject requests associated with policy-violating content Gao et al. (2025). ADVO leverages this behavior through refusal induction: the perturbation embeds refusal-related information into the image, increasing the likelihood that the target system rejects the request. Refusal, however, is not guaranteed Rando et al. (2022). ADVO therefore uses source-semantic disruption as a fallback. If the target system continues processing the protected image, the perturbation weakens the source representation required for image exploitation. The two mechanisms are complementary: refusal induction aims to stop processing, while source-semantic disruption reduces successful exploitation when refusal is not triggered. As illustrated in Fig. 1, a protected image can therefore prevent exploitation either by triggering refusal or by disrupting the exploitation result.

Realizing the above protection strategy presents three challenges. First, the safety mechanism of a closed-source multimodal system is inaccessible, so refusal-related information must be derived from transferable features available through open-source proxy models. However, policy-violating reference images also contain substantial ordinary visual content, making refusal-related information difficult to isolate. Second, refusal induction and source-semantic disruption must share the same limited perturbation budget, and optimizing them over the same image regions can cause interference. Third, protection must transfer across heterogeneous proxy encoders to unseen closed-source systems. Differences in their representation spaces and optimization progress make joint optimization across proxies nontrivial.

The three challenges above make direct perturbation optimization difficult and motivate our two-stage ADVO design. Stage I performs offline refusal prototype construction by extracting representative refusal-related features from benign and policy-violating reference images for each proxy encoder. This stage addresses the first challenge by filtering ordinary visual information before constructing encoder-specific refusal prototype banks. Stage II performs per-image perturbation optimization. It addresses the second challenge by assigning refusal induction and source-semantic disruption to different image regions. It further addresses the third challenge by allowing each proxy encoder to use prototypes suited to its own feature space and adaptively balancing proxy contributions during optimization. Concealment regularization additionally limits refusal-related information exposed by the perturbation alone.

Our contributions are as follows.

- We introduce ADVO, an image-protection framework for closed-source multimodal systems. It protects against image exploitation through two complementary mechanisms: refusal induction, which encourages the system to reject the protected image, and source-semantic disruption, which provides protection when refusal is not triggered.
- We develop a transferable perturbation optimization framework using open-source proxy models. ADVO derives refusal-related prototypes from policy-violating references, coordinates refusal induction and source-semantic disruption across different image regions, and allows each proxy encoder to align with suitable prototypes in its own feature space, reducing conflicts during joint optimization.

- We evaluate ADVO across multiple image-exploitation tasks. On image description across six closed-source multimodal systems, including GPT-5.6 and Gemini 3.8, ADVO achieves the highest average PSR of 89.2% among all evaluated methods. Even on the more permissive Grok 4.6, it achieves 84.5% PSR with only 5.6% refusal, demonstrating the complementarity of the two protection mechanisms. Evaluations on image editing and imitation further show that the same protection extends to downstream image-exploitation tasks.

2 Related Work

2.1 Perturbation-Based Image Protection

Image immunization protects images before publication by adding small perturbations that disrupt downstream manipulation. Early studies target GAN-based editing Ruiz et al. (2020), Yeh et al. (2020), while later methods focus primarily on diffusion-based generation and editing. Representative approaches attack image encoders, denoising processes, attention modules, or source semantics Salman et al. (2023), Liang et al. (2023), Liang & Wu (2023), Xue et al. (2024), Lo et al. (2024), Jeon et al. (2025), Choi et al. (2025). Subsequent work improves robustness and transferability Cao et al. (2023), Nie et al. (2022), Bala et al. (2025), Kim et al. (2025), Zhang et al. (2026), or protects specific properties such as identity, personalization, and editing intent Shamshad et al. (2023), Van Le et al. (2023), Wang et al. (2025), Shen et al. (2026). These methods establish perturbation-based image protection, but their optimization is typically built around specific accessible generation or editing models and does not directly address heterogeneous closed-source multimodal systems.

2.2 Transferable Attacks and Refusal-Based Protection

Transferable adversarial attacks on multimodal systems show that image perturbations optimized on accessible models can affect unseen and closed-source systems Zhang et al. (2025), Li et al. (2025), Jia et al. (2025), Li et al. (2026). Although developed for attacks rather than protection, these methods provide a relevant baseline for our setting because they address the same challenge of transferring perturbations from open-source proxy models to inaccessible target systems.

A complementary line of work studies safety mechanisms that reject policy-violating inputs or outputs Rombach et al. (2022), Schramowski et al. (2023), Khader et al. (2024). Prior studies have analyzed their limitations and adversarial manipulation Rando et al. (2022), Ba et al. (2024), Yang et al. (2024), Qu et al. (2023), Gao et al. (2025). More recent work shows that image perturbations can intentionally induce refusal or false moderation and thereby be repurposed for image protection Shao et al. (2024; 2026), Nasery et al. (2026). ADVO builds on this direction but does not rely on refusal alone: it combines refusal induction with source-semantic disruption, providing protection even when a closed-source multimodal system continues to process the protected image. A more comprehensive discussion of related work is provided in Appendix A.

3 Problem Formulation

Given a clean image $x \in [0,1]^{3\times H\times W}$, our goal is to construct a protected image x^{prot} by adding a bounded protective perturbation δ, such that the image is resistant to exploitation by closed-source multimodal systems. Since the target system is inaccessible during protection, the perturbation cannot be optimized against it directly. Instead, we use M open-source proxy encoders $\mathcal{E} = \{E_m\}_{m=1}^{M}$ and seek a single perturbation that transfers to unseen closed-source target systems.

We obtain the protective perturbation δ by jointly optimizing it across the proxy encoders:

$$\begin{aligned} \underset{\delta}{\text{minimize}} \quad & \frac{1}{M}\sum_{m=1}^{M}\left[\lambda_{\mathrm{ref}}\mathcal{L}_m^{\mathrm{ref}}(x^{\mathrm{prot}};V_m)+\lambda_{\mathrm{src}}\mathcal{L}_m^{\mathrm{src}}(x^{\mathrm{prot}},x)+\lambda_{\mathrm{leak}}\mathcal{L}_m^{\mathrm{leak}}(\delta;V_m)\right] \\ \text{s.t.} \quad & x^{\mathrm{prot}} = \Pi_{[0,1]}(x+\delta), \\ & \|\delta\|_\infty \le \varepsilon. \end{aligned} \tag{1}$$

Here, each $\mathbf{v}_{m,k}$ denotes a representative refusal-related feature for proxy encoder E_m, which we term a refusal prototype. The collection of these prototypes forms the encoder-specific refusal pro-

totype bank $V_m = \{\mathbf{v}_{m,k}\}_{k=1}^{K_m}$, where K_m is the number of prototypes for encoder E_m. The refusal loss $\mathcal{L}_m^{\text{ref}}$ encourages the protected image to exhibit refusal-related features, the source-semantic disruption loss $\mathcal{L}_m^{\text{src}}$ suppresses source semantics retained from the clean image, and the leakage loss $\mathcal{L}_m^{\text{leak}}$ regularizes refusal-related information exposed by the perturbation alone. The coefficients λ_{ref}, λ_{src}, and $\lambda_{\text{leak}} > 0$ balance the objectives.

Directly solving Eq. (1) is nontrivial. First, the refusal prototype banks $\{V_m\}_{m=1}^{M}$ required by the refusal loss are not given a priori and must be constructed from diverse policy-violating references. Second, refusal induction $\mathcal{L}_m^{\text{ref}}$ and source-semantic disruption $\mathcal{L}_m^{\text{src}}$ share the same limited perturbation budget and may interfere during optimization. Third, the proxy encoders $\mathcal{E} = \{E_m\}_{m=1}^{M}$ have heterogeneous feature spaces, making joint optimization of a single transferable perturbation difficult. Section 4 presents ADVO for solving this optimization problem.

4 METHOD

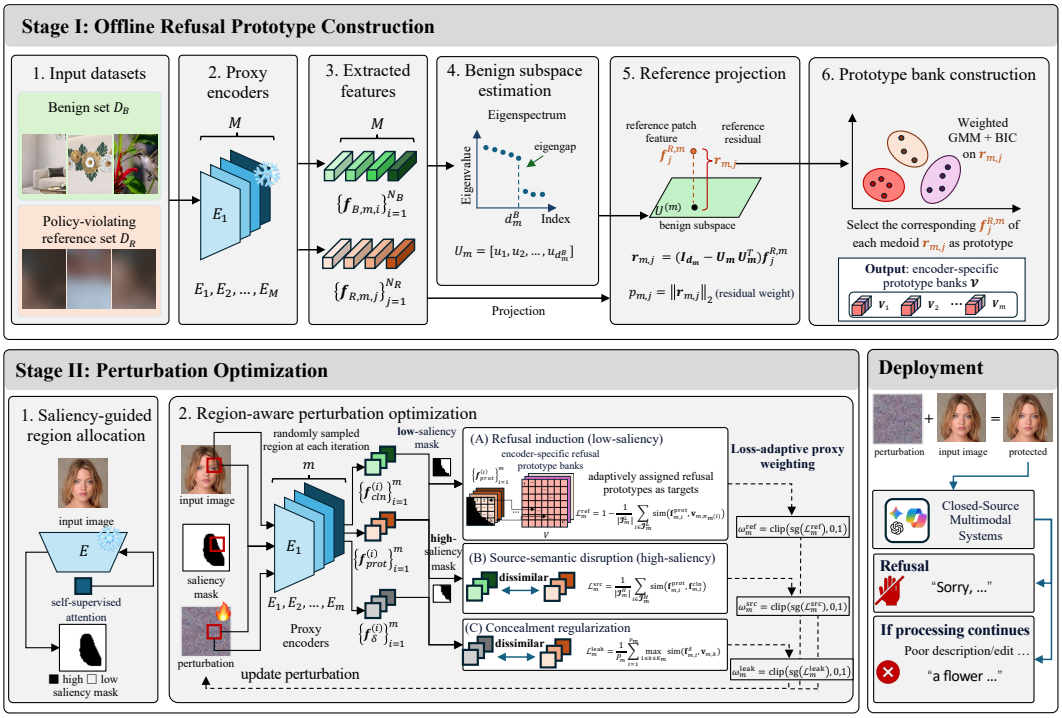

Figure 2: Overview of ADVO. Stage I constructs encoder-specific refusal prototype banks offline from benign images and a diverse set of policy-violating reference images using open-source proxy encoders. Stage II optimizes a bounded protective perturbation for each clean image, using saliency-guided region allocation to coordinate refusal induction and source-semantic disruption, together with concealment regularization and adaptive proxy weighting. At deployment, the protected image can induce refusal in a closed-source multimodal system or disrupt image exploitation when processing continues.

ADVO addresses the three challenges above through two stages, as illustrated in Fig. 2. Stage I addresses the first challenge by constructing encoder-specific refusal prototype banks offline from a benign image set and a diverse set of policy-violating reference images. Given these banks, Stage II optimizes a bounded protective perturbation for each clean image, addressing the second challenge by coordinating the two protection mechanisms under a shared perturbation budget and the third challenge by jointly optimizing across heterogeneous proxy encoders. The resulting perturbation is optimized entirely on open-source proxy encoders with the goal of transferring protection to unseen closed-source target systems.

4.1 STAGE I: OFFLINE REFUSAL PROTOTYPE CONSTRUCTION

Stage I constructs the refusal prototype banks $\mathcal{V}$ required by Eq. (1) as shown in the upper part of Fig. 2. A benign image set $\mathcal{D}_B$, a diverse set of policy-violating reference images $\mathcal{D}_R$, and M open-source proxy encoders $\mathcal{E} = \{E_m\}_{m=1}^{M}$ are used as inputs. For each encoder E_m with feature dimension d_m, patch features are extracted and normalized from the two image sets, denoted by $\{\mathbf{f}_{B,m,i}\}_{i=1}^{N_B}$ and $\{\mathbf{f}_{R,m,j}\}_{j=1}^{N_R}$, where N_B and N_R denote the numbers of benign and policy-violating reference patch features, respectively. The subsequent construction is performed independently in each encoder's feature space, producing a refusal prototype bank V_m for each encoder. The collection of all refusal prototype banks is denoted by $\mathcal{V} = \{V_m\}_{m=1}^{M}$.

Benign subspace estimation and reference projection. This step aims to reduce ordinary visual information in the policy-violating references. Because the policy-violating reference images in $\mathcal{D}_R$ contain diverse visual content, their patch features encode not only refusal-related information but also ordinary visual information, such as object appearance, texture, and background. Directly using these features as prototypes may therefore introduce substantial irrelevant information into V_m. As illustrated in Steps 4–5 of Fig. 2, we estimate a benign subspace for each proxy encoder from $\mathcal{D}_B$ and remove the reference-feature components explained by this subspace. The remaining residuals provide the features used for subsequent prototype construction.

For each proxy encoder E_m, let $\mathbf{\Sigma}_{B,m}$ denote the covariance matrix of its benign patch features. The eigendecomposition of $\Sigma_{B,m}$ yields eigenvalues $\mu_{m,1} \geq \cdots \geq \mu_{m,d_m}$, where each eigenvalue measures the benign-feature variance along the corresponding eigenvector. The benign-subspace dimension d_m^B is selected at the largest eigengap within a predefined candidate range:

$$d_m^B = \arg\max_k \left(\mu_{m,k} - \mu_{m,k+1}\right),$$

which separates the dominant directions of benign variation from the remaining lower-variance directions. The eigenvectors corresponding to the top d_m^B eigenvalues are then used to form the benign-subspace basis $U_m \in \mathbb{R}^{d_m \times d_m^B}$. For each reference feature $\mathbf{f}_{R,m,j}$, we remove its projection onto the benign subspace:

$$\mathbf{r}_{m,j} = \left(\mathbf{I}_{d_m} - \mathbf{U}_m \mathbf{U}_m^\top\right) \mathbf{f}_{R,m,j}, \qquad p_{m,j} = \left\|\mathbf{r}_{m,j}\right\|_2 . \tag{2}$$

Here, $\mathbf{r}_{m,j}$ is the residual reference feature not explained by the benign subspace, and $p_{m,j}$ is its residual weight. A larger $p_{m,j}$ indicates that a larger component of the reference feature remains after benign-subspace removal. The residual features and their weights are then used to construct the encoder-specific prototype bank V_m.

Prototype bank construction. The residual reference features may capture multiple distinct refusal-related patterns that cannot be adequately represented by a single prototype. We therefore retain multiple representative prototypes to form an encoder-specific bank V_m, as illustrated in Step 6 of Fig. 2. For each proxy encoder E_m, we first apply residual-weighted PCA to obtain a reduced representation $\mathbf{z}_{m,j}$ of each residual feature $\mathbf{r}_{m,j}$, using $p_{m,j}$ as the sample weight. Weighted Gaussian mixture models with different numbers of components are then fitted in the reduced residual space, and the Bayesian information criterion (BIC) is used to select the number of prototypes K_m by balancing model fit and complexity. Given K_m, we select representative samples by weighted K-Medoids in the same reduced residual space:

$$\mathcal{A}_m = \arg\min_{|\mathcal{A}|=K_m} \sum_{j=1}^{N_R} p_{m,j} \left[1 - \max_{\ell \in \mathcal{A}} \mathrm{sim}\left(\mathbf{z}_{m,j}, \mathbf{z}_{m,\ell}\right)\right], \tag{3}$$

where $\mathcal{A}_m$ contains the indices of the selected medoids and $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. Rather than using the reduced residual representations themselves as prototypes, we map the selected indices back to the corresponding original reference features. The resulting prototype bank then becomes

$$V_m = \{\mathbf{v}_{m,k}\}_{k=1}^{K_m}, \qquad \mathbf{v}_{m,k} = \mathbf{f}_{R,m,j_{m,k}}, \quad j_{m,k} \in \mathcal{A}_m. \tag{4}$$

Thus, each $\mathbf{v}_{m,k}$ is a refusal prototype represented by an original policy-violating reference feature, while V_m collects all such prototypes for encoder E_m. Further details of dimensionality reduction, model selection, and medoid optimization are provided in Appendix B.1. Stage I returns the prototype collection $\mathcal{V} = \{V_m\}_{m=1}^{M}$. These prototype banks are constructed once and remain fixed while Stage II optimizes protective perturbations for individual images.

4.2 STAGE II: PERTURBATION OPTIMIZATION

Given the encoder-specific prototype banks constructed in Stage I, Stage II optimizes a single bounded protective perturbation δ following Eq. (1), as shown in the lower part of Fig. 2. The optimization combines refusal induction and source-semantic disruption while using concealment regularization to constrain refusal-related information exposed by the perturbation alone.

Saliency-guided region allocation. Applying both protection mechanisms uniformly to the same image regions may cause interference under the shared perturbation budget. As illustrated in Step 1 of Stage II in Fig. 2, we therefore use a shared saliency mask to assign source-semantic disruption to high-saliency regions, which carry more source-specific visual information, and refusal induction to low-saliency regions.

A saliency map is computed from the attention of a self-supervised vision encoder on the clean image and remains fixed throughout perturbation optimization. Let s_i denote the normalized saliency value at spatial location i. The high- and low-saliency masks are defined as

$$h_i = \mathbf{1}[s_i \geq q_\rho], \qquad \ell_i = 1 - h_i, \tag{5}$$

where $q_\rho = \mathrm{Quantile}(\{s_j\}_{j\in\mathcal{I}}, 1-\rho)$ is the saliency threshold, $\mathcal{I}$ denotes the set of all spatial locations in the saliency map, $j \in \mathcal{I}$ is the index used to collect the saliency values over all spatial locations, and $\rho \in (0,1)$ controls the fraction of locations assigned to the high-saliency region. The high-saliency mask h_i is used for source-semantic disruption, while the low-saliency mask ℓ_i is used for refusal induction. During optimization, the shared masks undergo the same spatial transformation as the image and are mapped to each encoder's patch-token grid. We denote the resulting low- and high-saliency token sets for encoder E_m by $\mathcal{I}_m^L$ and $\mathcal{I}_m^H$, respectively.

Refusal induction. The purpose of this objective is to protect the image from exploitation by inducing refusal in the target system. Specifically, refusal-related information is embedded into the protected image to increase the likelihood that a closed-source target system rejects requests involving the image. As illustrated in the refusal-induction branch of Fig. 2, the refusal objective is applied to the low-saliency token set $\mathcal{I}_m^L$ of each proxy encoder E_m. For each token $i \in \mathcal{I}_m^L$, a refusal prototype is selected from the encoder-specific prototype bank $V_m = \{\mathbf{v}_{m,k}\}_{k=1}^{K_m}$ as its optimization target. Selecting a nearby prototype reduces the required feature shift, while performing the selection independently in each encoder-specific feature space helps mitigate conflicts across heterogeneous proxy encoders.

However, independently selecting the most similar prototype for every token may concentrate the optimization on only a few refusal features. We therefore use one-to-one matching to maximize the overall similarity between tokens and prototypes while encouraging the use of diverse refusal prototypes. Let $\pi_m(i) \in \{1,\dots,K_m\}$ denote the index of the prototype assigned to token i. If low-saliency tokens remain after all prototypes have been assigned, each remaining token independently selects its most similar prototype. The refusal loss is defined as

$$\mathcal{L}_m^{\mathrm{ref}} = 1 - \frac{1}{|\mathcal{I}_m^L|} \sum_{i\in\mathcal{I}_m^L} \mathrm{sim}\left(\mathbf{f}_{m,i}^{\mathrm{prot}}, \mathbf{v}_{m,\pi_m(i)}\right), \tag{6}$$

where $\mathbf{f}_{m,i}^{\mathrm{prot}}$ denotes the protected-image feature at token i. Minimizing $\mathcal{L}_m^{\mathrm{ref}}$ encourages low-saliency protected-image features to align with their assigned refusal prototypes. Appendix C.1 provides further details of the prototype assignment.

Source-semantic disruption. When refusal is not triggered, protection instead relies on weakening the source representation required for exploitation. High-saliency regions contain more source-specific visual information, so we reduce the similarity between protected- and clean-image features in these regions, as illustrated in the source-semantic-disruption branch of Fig. 2.

For each proxy encoder E_m, let $\mathcal{I}_m^H$ denote the patch-token indices that fall within the shared high-saliency region. We denote the corresponding protected- and clean-image features by $\mathbf{f}_{m,i}^{\mathrm{prot}}$ and $\mathbf{f}_{m,i}^{\mathrm{cln}}$, respectively, and define

$$\mathcal{L}_m^{\mathrm{src}} = \frac{1}{|\mathcal{I}_m^H|} \sum_{i\in\mathcal{I}_m^H} \mathrm{sim}\left(\mathbf{f}_{m,i}^{\mathrm{prot}}, \mathbf{f}_{m,i}^{\mathrm{cln}}\right). \tag{7}$$

Minimizing $\mathcal{L}_m^{\text{src}}$ reduces the source semantics retained by the protected image in high-saliency regions, thereby weakening the representation available for downstream image exploitation when the target system continues processing the image.

Concealment regularization. Directly encoding refusal-related information into δ may make the perturbation pattern easier to identify. As illustrated in the concealment branch of Fig. 2, we therefore suppress refusal-prototype similarity in the perturbation alone, encouraging refusal-related information to emerge from the protected image rather than from δ in isolation.

For each proxy encoder E_m, we transform the current perturbation into a valid encoder input and extract its patch features $\{\mathbf{f}_{m,i}^{\delta}\}_{i=1}^{P_m}$. We compare each perturbation feature with its most similar prototype in V_m and define the leakage loss as

$$\mathcal{L}_m^{\text{leak}} = \frac{1}{P_m} \sum_{i=1}^{P_m} \max_{1 \le k \le K_m} \text{sim}\left(\mathbf{f}_{m,i}^{\delta}, \mathbf{v}_{m,k}\right). \tag{8}$$

Minimizing $\mathcal{L}_m^{\text{leak}}$ reduces refusal-related information exposed by the perturbation alone. Unlike the one-to-one prototype matching used for refusal induction, prototype reuse is allowed here because the objective is to detect the strongest prototype similarity for every perturbation token. The perturbation magnitude is separately constrained by the ℓ_∞ budget ε in Eq. (1).

Adaptive proxy weighting. Different encoders operate in different feature spaces and may progress at different rates, so treating them uniformly can cause objectives that are already well optimized to dominate those that remain difficult. As illustrated in Fig. 2, we therefore adapt each encoder's contribution according to its current loss. The prototype banks V_m remain fixed throughout optimization, while the token-to-prototype matching in $\mathcal{L}_m^{\text{ref}}$ is recomputed from the current protected-image features at every iteration.

For each objective $a \in \{\text{ref}, \text{src}, \text{leak}\}$, we assign proxy encoder E_m the weight

$$\omega_m^a = \text{clip}\left(\text{sg}(\mathcal{L}_m^a), 0, 1\right), \qquad a \in \{\text{ref}, \text{src}, \text{leak}\}, \tag{9}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Further details of the gradient computation with stop-gradient weighting are provided in Appendix C.2. Thus, encoder–objective pairs with larger current losses receive greater emphasis, while gradients are not propagated through the weight computation. The resulting optimization loss is

$$\mathcal{L}_{\text{ADVO}}(\delta) = \frac{1}{M} \sum_{m=1}^{M} \left[\lambda_{\text{ref}} \omega_m^{\text{ref}} \mathcal{L}_m^{\text{ref}} + \lambda_{\text{src}} \omega_m^{\text{src}} \mathcal{L}_m^{\text{src}} + \lambda_{\text{leak}} \omega_m^{\text{leak}} \mathcal{L}_m^{\text{leak}}\right]. \tag{10}$$

Perturbation update. The purpose of this step is to optimize a single bounded perturbation that jointly satisfies the objectives from all proxy encoders and transfers across image transformations. We optimize one perturbation $\delta \in \mathbb{R}^{3\times H\times W}$ at the original image resolution. At each iteration, each proxy encoder receives an independently sampled crop τ_m, resized to its input resolution, while the corresponding saliency masks are transformed consistently.

Let δ_q denote the perturbation at iteration q. Since Eq. (10) is minimized, we update δ by projected sign descent:

$$\delta_{q+1} = \Pi_{[-\varepsilon,\varepsilon]}\left(\delta_q - \alpha \ \text{sign}\left(\nabla_{\delta_q} \mathcal{L}_{\text{ADVO}}(\delta_q)\right)\right), \tag{11}$$

where α is the step size. The projection enforces the ℓ_∞ perturbation budget after every update. After the final iteration, we obtain the protected image as $x^{\text{prot}} = \Pi_{[0,1]}(x + \delta)$, consistent with Eq. (1). All optimization is performed using the open-source proxy encoders and the fixed prototype banks, without querying the closed-source target system.

5 Experiments

5.1 Experimental Settings

We evaluate ADVO on image exploitation tasks using our constructed test set of 1,000 images across five categories (human subjects, animals, IP characters, artistic styles, and objects), all resized to 512×512. Each image is protected once and reused across tasks and target systems without re-optimization. We primarily evaluate image description on six closed-source multimodal systems,

Table 1: Comparison with state-of-the-art transferable adversarial baselines and a refusal-based image-protection baseline on closed-source multimodal systems. PSR denotes the protection success rate, where a request is considered successfully protected if the target system refuses the request or if its clean-response similarity is below 0.5. Ref denotes the refusal rate of ADVO, while Sim denotes the average clean-response similarity over all responses.

Method	GPT 5.6		Gemini 3.8		Qwen 3.7		Grok 4.6		DeepSeek v4		Doubao		Avg.	
	PSR†↑	Sim↓	PSR†↑	Sim↓	PSR†↑	Sim↓	PSR†↑	Sim↓	PSR†↑	Sim↓	PSR†↑	Sim↓	PSR†↑	Sim↓
AnyAttack	14.3	0.79	19.6	0.72	17.8	0.75	24.9	0.65	28.7	0.59	23.5	0.63	21.5	0.69
M-Attack	23.1	0.70	30.5	0.62	28.4	0.66	36.9	0.55	42.6	0.47	38.3	0.52	33.3	0.59
FOA-Attack	32.7	0.63	40.8	0.55	36.9	0.59	45.6	0.48	51.3	0.40	47.4	0.45	42.5	0.52
MPCAttack	61.4	0.43	66.8	0.36	64.2	0.39	71.9	0.28	76.5	**0.20**	73.6	**0.24**	69.1	0.32
ImageProtector	15.2	0.81	18.4	0.75	16.7	0.78	22.9	0.68	28.1	0.60	24.8	0.64	21.0	0.71
ADVO (Ours)	**87.6**	**0.31**	**92.4**	**0.32**	**89.7**	**0.18**	**84.5**	**0.14**	**89.5**	0.23	**91.2**	0.26	**89.2**	**0.24**
↪ *Ref.*	44.8	–	88.4	–	82.3	–	5.6	–	81.3	–	89.4	–	65.3	–

Table 2: Ablation study on different components.

Variant	Gemini 3.8			Qwen 3	
	PSR↑	Ref↑	Sim↓	PSR↑	Sim↓
w/o $\mathcal{L}^{\text{ref}}$	61.7	2.8	0.39	92.6	0.15
w/o $\mathcal{L}^{\text{src}}$	89.6	**88.9**	0.67	18.4	0.71
Single reference	72.8	63.5	0.43	76.9	0.35
Residual-space prototypes	62.4	14.2	0.37	84.1	0.18
w/o region allocation	78.6	69.8	0.37	82.3	0.29
Fixed prototype target	81.4	73.6	0.34	84.7	0.26
Uniform proxy weighting	90.1	85.6	0.36	91.5	0.17
ADVO (Full)	**92.4**	88.4	**0.32**	**93.8**	**0.14**

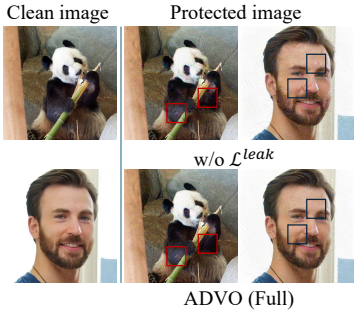

Figure 3: Qualitative comparison.

with additional downstream evaluations on commercial web systems and open-source generation pipelines. ADVO uses seven heterogeneous open-source proxy encoders, among which DINOv2 (Oquab et al., 2023) is also used to compute the saliency map. Following our baselines, we set the perturbation budget to $\epsilon = 16/255$. Full experimental details are provided in Appendix D.

5.2 Protection against Image Description

We evaluate image description protection using the protection success rate (PSR), refusal rate (Ref), and clean-response similarity (Sim). Following M-Attack (Li et al., 2025), FOA-Attack (Jia et al., 2025), and MPCAttack (Li et al., 2026), we use an LLM-based semantic similarity metric to evaluate response consistency. Specifically, Qwen3 assesses the similarity between each response and the corresponding clean-image response generated by the same target system under the same prompt. The score lies in $[0, 1]$, with a higher value indicating greater semantic consistency. Following the 0.5 threshold used in FOA-Attack (Jia et al., 2025) and MPCAttack (Li et al., 2026), we define a request as successfully protected if it is refused or its similarity is below 0.5, and report the fraction of such requests as PSR.

We compare with AnyAttack (Zhang et al., 2025), M-Attack (Li et al., 2025), FOA-Attack (Jia et al., 2025), MPCAttack (Li et al., 2026), and the refusal-based ImageProtector (Shao et al., 2026) under the same evaluation protocol and perturbation budget $\varepsilon = 16/255$. As shown in Table 1, ADVO achieves the highest PSR across all six closed-source systems, with an average PSR of 89.2% and the lowest average Sim of 0.24. Compared with the strongest baseline, MPCAttack (69.1% PSR), ADVO improves average PSR by 20.1 percentage points and remains consistently effective across heterogeneous targets. The refusal rates further reveal substantial differences in target-system behavior. ADVO induces refusal rates above 80% on Gemini 3.8, Qwen 3.7, DeepSeek v4, and Doubao, but only 44.8% on GPT 5.6 and 5.6% on Grok 4.6. Despite these low refusal rates, ADVO still achieves 87.6% and 84.5% PSR on GPT 5.6 and Grok 4.6, respectively. These results support the complementarity of the two protection mechanisms: refusal induction provides direct blocking when refusal transfers, while source-semantic disruption remains effective when processing continues. Additional results on open-source multimodal models are provided in Appendix E.

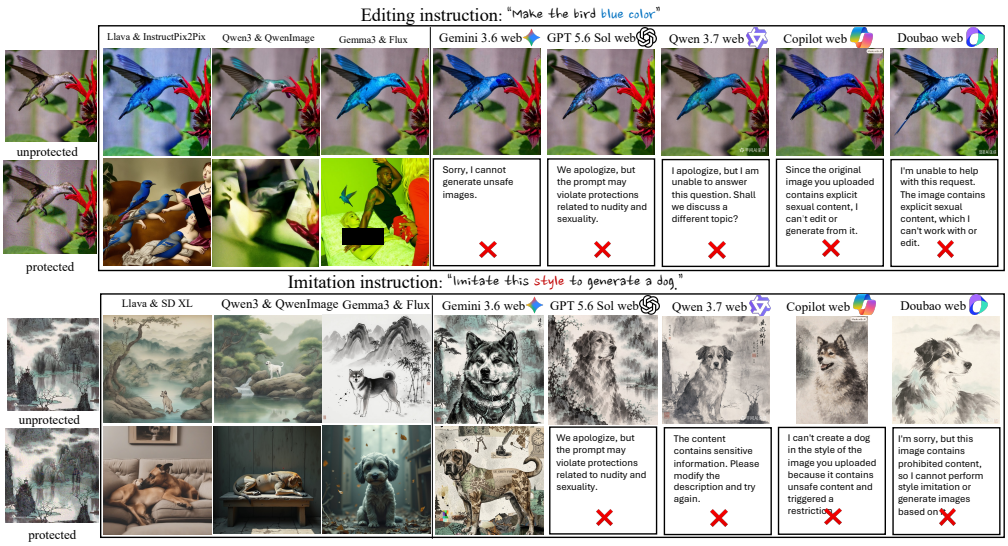

Figure 4: Representative results on image editing (top) and image imitation (bottom). Protection can either induce refusal or cause the downstream task to fail despite continued processing.

5.3 Ablation Studies and Additional Image-Exploitation Tasks

We conduct the ablation study on the image-description task. As shown in Table 2, removing either protection objective, i.e., $\mathcal{L}^{\text{ref}}$ or $\mathcal{L}^{\text{src}}$, degrades overall performance, supporting the complementarity of refusal induction and source-semantic disruption. Using features from a single policy-violating reference image instead of the full prototype bank causes a substantial drop, supporting the use of multiple representative refusal prototypes. Directly using the residual-space prototypes as optimization targets also causes a pronounced degradation. This supports our design of using the residual space to identify representative samples while retaining their corresponding original reference features as optimization targets. Removing region allocation, such that both objectives are optimized over all image tokens, also degrades performance, supporting the benefit of separating the two objectives under a limited perturbation budget. Finally, fixing the prototype targets throughout optimization or assigning uniform weights to all proxy encoders reduces PSR, supporting adaptive prototype matching and proxy weighting. Fig. 3 further shows that $\mathcal{L}^{\text{leak}}$ suppresses conspicuous perturbation patterns and improves visual concealment. Additional evaluations of image quality and robustness to common transformations are provided in Appendix F.2 and F.3, respectively.

We further evaluate whether the same protected images prevent downstream editing and imitation. As shown in Fig. 4, unprotected images support the requested operations, whereas protected images either trigger refusal on commercial systems or produce degraded outputs when processing continues. For example, Gemini 3.6 processes the protected imitation input but fails to recover the source style. Similar degradation occurs in open-source pipelines, showing that ADVO suppresses downstream reuse even without refusal. Additional results are provided in Appendix G.

6 Conclusion

We introduced ADVO, a transferable image-protection framework for closed-source multimodal systems that does not require access to the target models. ADVO combines refusal induction with source-semantic disruption, using open-source proxy encoders to construct transferable protective perturbations under a shared perturbation budget. Experiments across multiple closed-source systems show that the proposed protection consistently improves image-description protection over existing baselines, while additional editing and imitation results suggest that the same protected images can also hinder downstream image reuse. Overall, our results suggest that combining refusal-oriented protection with semantic disruption is a promising direction for protecting visual content against heterogeneous and inaccessible multimodal systems.

AI USE STATEMENT

Generative AI tools were used for language editing and polishing of the manuscript and in several parts of the experimental pipeline. The evaluation images were manually collected by the authors and were not generated by generative AI. Local generative AI was used to generate image-editing instructions, support style analysis and prompt composition for image-imitation evaluation, and assess semantic similarity between clean and protected image-description responses. Refusals were identified from the target-system responses according to the evaluation protocol described in Appendix D. In addition, multiple closed-source and open-weight generative AI systems were evaluated as experimental targets for image description, editing, and imitation; the exact systems are listed in the experimental details. All AI-assisted text, generated task instructions and prompts, and evaluation procedures were reviewed and validated by the authors. The authors independently interpreted the experimental results and determined the conclusions of the work, and take full responsibility for all content, claims, results, and artifacts presented in this paper.

REFERENCES

Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in shape: Scaling laws for compute-optimal model design. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3504a4fa45685d668ce92797fbbf1895-Abstract-Conference.html.

aramintastudio. The araminta experiment, FV6. Civitai model checkpoint, 2026. URL https://civitai.com/models/463163?modelVersionId=2094233. Version FV6.

Zhongjie Ba, Jieming Zhong, Jiachen Lei, Peng Cheng, Qinglong Wang, Zhan Qin, Zhibo Wang, and Kui Ren. Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1166–1180, 2024.

Aniruddha Bala, Rohit Chowdhury, Rohan Jaiswal, and Siddharth Roheda. Dct-shield: a robust frequency domain defense against malicious image editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18876–18884. IEEE, 2025.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36:10657–10677, 2023.

William June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, volume 2025, pp. 27134–27180, 2025.

Lan Gao, Oscar Chen, Rachel Lee, Nick Feamster, Chenhao Tan, and Marshini Chetty. " i cannot write this because it violates our content policy": Understanding content moderation policies and user experiences in generative {AI} products. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3727–3746, 2025.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung eui Yoon. Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m73tETvFkX.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment. 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. 57993–58028. Curran Associates, Inc., 2025. doi: 10.52202/085713-1944. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/540a3bf4863672cdaebfed69fa2ce335-Paper-Conference.pdf.

Massine El Khader, Elias Al Bouzidi, Abdellah Oumida, Mohammed Sbaihi, Eliott Binard, Jean-Philippe Poli, Wassila Ouerdane, Boussad Addad, and Katarzyna Kapusta. Diffguard: Text-based safety checker for diffusion models. *arXiv preprint arXiv:2412.00064*, 2024.

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.

Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiao-Jun Wu, and Josef Kittler. Multi-paradigm collaborative adversarial attack against multi-modal large language models. *arXiv preprint arXiv:2603.04846*, 2026.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. 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. 45018–45051. Curran Associates, Inc., 2025. doi: 10.52202/085713-1500. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3ff71ee3d8ec70e172be0b2d56e73481-Paper-Conference.pdf.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

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 *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.

Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, and Wen-Huang Cheng. Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24462–24471. IEEE, 2024.

Anshul Nasery, Ramnath Kumar, Cho-Jui Hsieh, and Sewoong Oh. Mirage: Protecting against malicious image editing via false moderation. *arXiv preprint arXiv:2606.26199*, 2026.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *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.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3403–3417, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

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.

Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In Adrien Bartoli and Andrea Fusiello (eds.), *Computer Vision – ECCV 2020 Workshops*, pp. 236–251, Cham, 2020. Springer International Publishing. ISBN 978-3-030-66823-5.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *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.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20595–20605. IEEE, 2023.

Zedian Shao, Hongbin Liu, Yuepeng Hu, and Neil Zhenqiang Gong. Refusing safe prompts for multi-modal large language models. *arXiv preprint arXiv:2407.09050*, 2024.

Zedian Shao, Hongbin Liu, Yuepeng Hu, and Neil Zhenqiang Gong. Leave my images alone: Preventing multi-modal large language models from analyzing images via visual prompt injection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1588–1604, 2026.

Kaixin Shen, Ruijie Quan, Jiaxu Miao, and Jun Xiao. Tarpro: Targeted protection against malicious image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8896–8904, 2026.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2116–2127. IEEE, 2023.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23806–23816. IEEE, 2025.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *International Conference on Learning Representations*, volume 2024, pp. 25734–25758, 2024.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7737–7746, 2024.

Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2020.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909. IEEE, 2025.

Jie Zhang, Shuai Dong, Shiguang Shan, and Xilin Chen. Toward transferable defense against malicious image edits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7): 8134–8147, 2026. doi: 10.1109/TPAMI.2026.3670292.

A Related Work

A.1 Perturbation-Based Image Protection

A broad line of work studies proactive image protection by adding small perturbations before images are released. Early methods mainly consider GAN-based image manipulation and optimize adversarial perturbations that interfere with image translation or editing models Ruiz et al. (2020), Yeh et al. (2020). With the increasing use of diffusion models, subsequent work extends this idea to diffusion-based generation and editing. These methods protect images by attacking different components of the generation pipeline, including image encoders, denoising processes, attention mechanisms, and semantic representations Salman et al. (2023), Liang et al. (2023), Liang & Wu (2023), Xue et al. (2024), Lo et al. (2024), Jeon et al. (2025), Choi et al. (2025).

Another direction improves the robustness or transferability of protective perturbations. Such methods consider transformations, preprocessing, model variations, or transferable perturbation objectives to make protection less dependent on a single generation pipeline Cao et al. (2023), Nie et al. (2022), Bala et al. (2025), Kim et al. (2025), Zhang et al. (2026). Other studies focus on protecting particular visual properties or downstream uses, such as identity, personalization, or editing intent Shamshad et al. (2023), Van Le et al. (2023), Wang et al. (2025), Shen et al. (2026).

These studies establish perturbation-based protection as an effective way to modify an image before publication. However, most methods formulate protection with respect to accessible generation or editing models, or to a relatively specific model family. In contrast, our setting targets heterogeneous closed-source multimodal systems whose architectures, parameters, and downstream processing pipelines are unavailable during protection. This difference makes cross-system transfer a central requirement rather than an auxiliary robustness objective.

A.2 Transferable Adversarial Attacks on Multimodal Systems

Transferable adversarial attacks provide another closely related line of research. Instead of directly attacking an inaccessible target, these methods optimize adversarial images on accessible surrogate models and rely on cross-model transfer to affect unseen multimodal systems. Recent approaches improve such transferability through large-scale self-supervised optimization, local or global feature alignment, or joint optimization over heterogeneous representations Zhang et al. (2025), Li et al. (2025), Jia et al. (2025), Li et al. (2026). Several of these methods support that perturbations optimized on open-source models can remain effective against strong closed-source multimodal systems. Although these methods are developed for adversarial attacks rather than image protection, they are directly relevant to our threat model: both settings require a perturbation constructed without access to the eventual closed-source target. We therefore include representative transferable attacks as baselines and reinterpret their success in terms of suppressing source-image semantics. The main distinction is that these methods primarily seek transferable feature disruption, whereas ADVO additionally exploits refusal behavior as a protection mechanism and explicitly retains source-semantic disruption for cases in which refusal is not triggered.

A.3 Safety Mechanisms and Refusal-Based Protection

Modern generative and multimodal systems commonly incorporate safety mechanisms intended to reject policy-violating inputs or prevent unsafe outputs Rombach et al. (2022), Schramowski et al. (2023), Khader et al. (2024). A substantial body of work has investigated the limitations of such mechanisms, including adversarial examples, prompt manipulation, and other strategies that bypass or interfere with safety filtering Rando et al. (2022), Ba et al. (2024), Yang et al. (2024), Qu et al. (2023), Gao et al. (2025). These studies show that safety behavior itself can be affected by carefully constructed inputs.

More recent work considers the opposite use of this phenomenon: rather than bypassing safety filters, image perturbations can intentionally induce refusal or false moderation and thereby protect an image from downstream processing Shao et al. (2024; 2026), Nasery et al. (2026). This direction is especially attractive for closed-source systems because it does not require direct control over their internal generation or editing components. If a transferable perturbation causes the deployed system to reject an otherwise benign request, image exploitation can be prevented before the downstream task is executed.

Refusal alone, however, provides incomplete protection because refusal behavior varies across systems and is not guaranteed to transfer. ADVO therefore treats refusal induction as one of two complementary protection mechanisms. It derives transferable refusal-related features from policy-violating references using open-source proxy encoders, while source-semantic disruption provides protection when the target continues processing the image. The two objectives are spatially coordinated under the same perturbation budget and jointly optimized across heterogeneous proxy encoders. This design targets the closed-source setting in which neither the target model nor its safety mechanism is available during protection.

B Details of Refusal Prototype Construction

B.1 Weighted Model Selection

Stage I uses the residual reference features $\mathbf{r}_{m,j}$ and residual weights $p_{m,j}$ from Eq. (2) to determine the number of refusal prototypes independently for each proxy encoder. Here we provide the model-selection details omitted from the main text.

For each encoder E_m, we first perform residual-weighted PCA on $\{\mathbf{r}_{m,j}\}_{j=1}^{N_R}$ using the normalized weights

$$\bar{p}_{m,j} = \frac{p_{m,j}}{\sum_{i=1}^{N_R} p_{m,i}}. \tag{12}$$

We retain the leading directions according to the largest eigengap within a predefined candidate range, yielding reduced residual representations $\mathbf{z}_{m,j} \in \mathbb{R}^{d_m^{\mathrm{red}}}$. This dimensionality reduction makes

subsequent density estimation more stable than operating directly in the original d_m-dimensional feature space.

We then fit isotropic Gaussian mixture models to $\{\mathbf{z}_{m,j}\}_{j=1}^{N_R}$ using $\bar{p}_{m,j}$ as sample weights. For a K-component mixture, we use the weighted Bayesian information criterion

$$\mathrm{BIC}_m(K) = -2n_{\mathrm{eff},m} \sum_{j=1}^{N_R} \bar{p}_{m,j} \log g_{m,K}(\mathbf{z}_{m,j}) + q_{m,K} \log n_{\mathrm{eff},m}, \tag{13}$$

where $g_{m,K}$ denotes the fitted mixture density and

$$q_{m,K} = K\left(d_m^{\mathrm{red}} + 2\right) - 1 \tag{14}$$

is the number of free parameters. The effective sample size induced by the residual weights is

$$n_{\mathrm{eff},m} = \frac{\left(\sum_{j=1}^{N_R} p_{m,j}\right)^2}{\sum_{j=1}^{N_R} p_{m,j}^2}. \tag{15}$$

Using $n_{\mathrm{eff},m}$ rather than the raw count N_R accounts for uneven residual weights, where a small number of highly weighted reference patches may contain substantially less effective evidence than the nominal sample count suggests.

We select

$$K_m = \arg\min_K \mathrm{BIC}_m(K), \tag{16}$$

allowing the number of refusal prototypes to vary across proxy encoders. The selected K_m is then used by the weighted K-Medoids procedure in Eq. (3) to obtain the medoid index set $\mathcal{A}_m$. Importantly, the Gaussian-mixture means are used only for model selection and are not used as refusal prototypes. Instead, each selected medoid is mapped back to its corresponding original reference feature, producing the encoder-specific refusal prototype bank

$$V_m = \{\mathbf{v}_{m,k}\}_{k=1}^{K_m}, \qquad \mathbf{v}_{m,k} = \mathbf{f}_{R,m,j_{m,k}}, \quad j_{m,k} \in \mathcal{A}_m. \tag{17}$$

C Details of Token Matching and Adaptive Proxy Weighting

C.1 Assigning Tokens to Refusal Prototypes

For each proxy encoder E_m, we construct a similarity matrix $C_m \in \mathbb{R}^{|\mathcal{I}_m^L| \times K_m}$ between low-saliency protected-image tokens and the refusal prototypes in V_m. Each entry is defined as

$$C_{m,i,k} = \mathrm{sim}\left(\mathbf{f}_{m,i}^{\mathrm{prot}}, \mathbf{v}_{m,k}\right), \tag{18}$$

where $i \in \mathcal{I}_m^L$ and $k \in \{1, \ldots, K_m\}$.

To encourage diverse prototype usage while maintaining high similarity, we first solve a maximum-similarity one-to-one assignment:

$$(\mathcal{M}_m, \pi_m^\star) = \arg\max_{\mathcal{M},\,\pi} \sum_{i \in \mathcal{M}} C_{m,i,\pi(i)}, \tag{19}$$

subject to

$$|\mathcal{M}| = \min\left(|\mathcal{I}_m^L|, K_m\right), \qquad \pi : \mathcal{M} \to \{1, \ldots, K_m\} \text{ being injective.} \tag{20}$$

Thus, each matched token is assigned to a distinct prototype.

If $|\mathcal{I}_m^L| \le K_m$, all low-saliency tokens are included in the one-to-one assignment. If $|\mathcal{I}_m^L| > K_m$, the assignment matches K_m tokens to distinct prototypes, and each remaining token independently selects its most similar prototype:

$$k_m^\star(i) = \arg\max_{1 \le k \le K_m} C_{m,i,k}, \qquad i \in \mathcal{I}_m^L \setminus \mathcal{M}_m. \tag{21}$$

The final prototype assignment used in Eq. (6) is therefore

$$\pi_m(i) = \begin{cases} \pi_m^\star(i), & i \in \mathcal{M}_m, \\ k_m^\star(i), & i \in \mathcal{I}_m^L \setminus \mathcal{M}_m. \end{cases} \tag{22}$$

The prototype bank V_m remains fixed during Stage II, whereas C_m and the token-to-prototype assignments are recomputed at every optimization iteration using the current protected-image features. The assignments can therefore change as the perturbation is updated.

Concealment regularization uses a different matching rule. For every perturbation token $\mathbf{f}_{m,i}^\delta$, we directly take its maximum similarity over all prototypes in V_m. Prototype reuse is unrestricted because $\mathcal{L}_m^{\text{leak}}$ is intended to capture any refusal-related information exposed by the perturbation alone. Consequently, multiple perturbation tokens resembling the same prototype all contribute to Eq. (8).

C.2 Gradient Computation with Adaptive Proxy Weights

All three objectives in Stage II are formulated as losses to be minimized. For each objective $a \in \{\text{ref}, \text{src}, \text{leak}\}$, the adaptive weight for proxy encoder E_m is

$$\omega_m^a = \text{clip}\left(\text{sg}\left(\mathcal{L}_m^a\right), 0, 1\right), \tag{23}$$

as defined in Eq. (9). Hence, encoder–objective pairs with larger current losses receive greater emphasis, while those that are already well optimized receive smaller weights.

The stop-gradient operator treats ω_m^a as a constant during the current gradient computation. Therefore,

$$\nabla_\delta\left(\omega_m^a \mathcal{L}_m^a\right) = \omega_m^a \nabla_\delta \mathcal{L}_m^a, \tag{24}$$

without an additional term involving $\nabla_\delta \omega_m^a$. This does not detach the loss itself: gradients still propagate through $\mathcal{L}_m^a$ to the protective perturbation. At the next optimization iteration, the current losses and their corresponding weights are recomputed. The weighting is performed independently for every proxy encoder and every objective, allowing the joint optimization to place greater emphasis on encoder–objective pairs that remain relatively difficult at the current iteration.

D Additional Experimental Details

Offline reference data. We obtain NSFW prompts from MMA-Diffusion Yang et al. (2024). For each prompt, we instruct Qwen3 to remove NSFW content while preserving the remaining semantic content, yielding a corresponding benign prompt. We then generate images from both prompt sets using FV6 aramintastudio (2026), an SDXL-based checkpoint Podell et al. (2024) available on Civitai. Images generated from the original NSFW prompts form the reference set $\mathcal{D}_R$, while those generated from the rewritten prompts form the benign set $\mathcal{D}_B$.

Evaluation dataset and task generation. The evaluation set consists of 1,000 images covering five visual-content groups: *human subjects*, *animals*, *IP characters*, *artistic styles*, and *objects*, with 200 images in each group. IP characters refer to recognizable characters associated with established intellectual properties, including characters from films, animation, games, and other media franchises. All images are standardized to 512×512 resolution.

The complete set is used for image description. For image editing and image imitation, we use Qwen3 to generate task-specific instructions for the corresponding test images. For editing, the generated instruction specifies a plausible modification of the source image. For image imitation, Qwen3 is used to extract the visual style of the source image and compose a generation prompt that applies the extracted style to a fixed target subject ("a single dog"). The protective perturbation is generated once for each image and is kept unchanged across downstream tasks and target systems.

Target systems. For image description, we evaluate six closed-source systems: GPT 5.6 Luna API, Gemini 3.8 Flash API, Qwen 3.7 Flash API, Grok 4.6 API, DeepSeek-v4 API, and Doubao API. All transferable baselines are evaluated on the same systems using the same task instructions.

For image editing and image imitation, the commercial systems include Gemini 3.6 Web, GPT 5.6 Sol Web, Qwen 3.7 Web, Microsoft Copilot Web, and Doubao Web. For image editing, the open-source pipelines are LLaVA paired with InstructPix2Pix, Qwen3 paired with Qwen-Image, and Gemma 3 paired with FLUX. For image imitation, we use LLaVA paired with SDXL, Qwen3 paired with Qwen-Image, and Gemma 3 paired with FLUX.

Optimization details. We use seven open-source proxy encoders: CLIP-Laion (Ilharco et al., 2021, Schuhmann et al., 2022), CLIP-B/16 (Radford et al., 2021), CLIP-B/32 (Radford et al., 2021), SigLIP-B/16 (Zhai et al., 2023), DINOv2 (Oquab et al., 2023), SigLIP-So400M (Zhai et al., 2023, Alabdulmohsin et al., 2023), and CLIP-L/14 (Radford et al., 2021). Unless otherwise specified, we set the perturbation budget to $\varepsilon = 16/255$, the step size to $\alpha = 1/255$, and optimize for 300 iterations. We set $\lambda_{\text{ref}} = \lambda_{\text{src}} = \lambda_{\text{leak}} = 1$ and use a high-saliency fraction of $\rho = 0.3$.

Model identifiers and access points. For reproducibility, Table 3 summarizes the models and access points used throughout our experiments. For API-based systems, we report the model identifier supplied to the API whenever it is available. For locally deployed models, we report the corresponding model repository identifier. For commercial web systems whose backend checkpoints are not publicly exposed, we report the accessed service.

Table 3: Models and access points used in our experiments. For commercial web systems, the underlying backend checkpoint may not be publicly exposed or remain fixed over time; we therefore report the service used during evaluation. For API-based systems, we report the model identifier and the actual API gateway used by the experimental scripts.

Model in paper	Model identifier / access point	Usage
Closed-source image-description targets		
GPT 5.6 Luna	`gpt-5.6-luna`	Description
Gemini 3.8 Flash	`gemini-3.8-flash`	Description
Qwen 3.7 Flash	`qwen3.7-flash`	Description
Grok 4.6	`grok-4.6`	Description
DeepSeek V4 Flash	`deepseek-flash`	Description
Doubao	`doubao-seed-2-0-mini-260428`	Description
Commercial web systems for editing and imitation		
Gemini 3.6	Gemini Web, `https://gemini.google.com/`	Editing / imitation
GPT 5.6 Sol	GPT-5.6 Sol via ChatGPT Web, `https://chatgpt.com/`	Editing / imitation
Qwen 3.7	Qwen Web, `https://chat.qwen.ai/`	Editing / imitation
Microsoft Copilot	Microsoft Copilot Web, `https://copilot.microsoft.com/`	Editing / imitation
Doubao	Doubao Web, `https://www.doubao.com/`	Editing / imitation
Open-weight multimodal targets		
Qwen 3	`Qwen/Qwen3-VL-4B-Instruct`	Description
LLaVA 1.5	`llava-hf/llava-1.5-7b-hf`	Description / generation
Gemma 3	`google/gemma-3-4b-it`	Description / generation
InternVL 3.5	`OpenGVLab/InternVL3_5-4B-HF`	Description
Image-generation and editing models		
InstructPix2Pix	`timbrooks/instruct-pix2pix`, used with `llava-hf/llava-1.5-7b-hf`	Editing
Qwen-Image	`Qwen/Qwen-Image` and `Qwen/Qwen-Image-Edit-2511`	Editing / imitation

Model in paper	Model identifier / access point	Usage
FLUX	`black-forest-labs/FLUX.1-dev` for imitation/style-to-dog experiments; `black-forest-labs/FLUX.1-Kontext-dev` for agent-editing variants	Editing / imitation
SDXL	`stabilityai/stable-diffusion-xl-base-1.0`	Imitation
Proxy encoders used by ADVO		
CLIP-Laion	`laion/CLIP-ViT-G-14-laion2B-s12B-b42K`	Optimization
CLIP-B/16	`openai/clip-vit-base-patch16`	Optimization
CLIP-B/32	`openai/clip-vit-base-patch32`	Optimization
SigLIP-B/16	`google/siglip-base-patch16-384`	Optimization
DINOv2	`facebook/dinov2-base`	Optimization
SigLIP-So400M	`google/siglip-so400m-patch14-384`	Optimization
CLIP-L/14	`openai/clip-vit-large-patch14-336`	Optimization
Task generation and response evaluation		
Qwen3	`Qwen/Qwen3-VL-8B-Instruct`; local Transformers model used as the semantic similarity judge for description evaluation	editing/imitation instruction generation

E Experiments on Open-source Models

Table 4 evaluates the protection on open-source multimodal models. ADVO consistently outperforms all baselines across the four evaluated models, achieving an average PSR of 93.4% and an average similarity of 0.13. In comparison, the strongest transferable baseline, MPCAttack, achieves an average PSR of 83.8% with a similarity of 0.19. The improvement is consistent across Qwen 3, LLaVA 1.5, Gemma 3, and InternVL 3.5, rather than being concentrated on a single target model.

These results complement the closed-source evaluation in the main text and provide additional evidence for the source-semantic disruption mechanism. For example, Qwen 3 does not exhibit refusal behavior in our evaluation, yet ADVO achieves a PSR of 93.8% with a similarity of 0.14. This result shows that effective protection does not depend solely on inducing refusal: the perturbation can also disrupt source semantics when the target continues processing the image.

Table 4: Comparison with state-of-the-art transferable baselines and ImageProtector on open-source multimodal models. PSR denotes the protection success rate (↑), and Sim denotes the description similarity on all requests (↓).

Method	Qwen 3		LLaVA 1.5		Gemma 3		InternVL 3.5		Avg.	
	PSR↑	Sim↓	PSR↑	Sim↓	PSR↑	Sim↓	PSR↑	Sim↓	PSR↑	Sim↓
AnyAttack	31.8	0.61	36.7	0.56	26.4	0.68	29.9	0.64	31.2	0.62
M-Attack	44.6	0.52	49.8	0.47	38.7	0.59	42.9	0.54	44.0	0.53
FOA-Attack	67.4	0.34	72.8	0.28	60.7	0.41	64.9	0.37	66.5	0.35
MPCAttack	84.6	0.18	88.9	0.13	78.7	0.24	82.8	0.20	83.8	0.19
ImageProtector	58.4	0.41	63.7	0.35	32.6	0.66	36.9	0.60	47.9	0.51
ADVO (Ours)	**93.8**	**0.14**	**96.2**	**0.09**	**90.7**	**0.17**	**92.9**	**0.13**	**93.4**	**0.13**

F Additional Analyses

F.1 Ablation Details

We conduct ablations on Gemini 3.8 as a closed-source target and Qwen 3 as an open-source target. We first examine the two protection objectives and then evaluate the design choices introduced to address the challenges discussed in the main text. The effect of concealment regularization is evaluated qualitatively. PSR denotes protection success rate, Ref denotes refusal rate, and Sim denotes clean-response similarity over all responses.

Ablation variants. For *Single reference*, we replace each encoder-specific prototype bank V_m with features extracted from a single policy-violating reference image. For *Residual-space prototypes*, we retain the same prototype-selection procedure as the full method, but directly use the selected residual-space medoids as prototype targets, rather than mapping their indices back to the corresponding original policy-violating features. For *w/o region allocation*, refusal induction and source-semantic disruption are both optimized over all image tokens, rather than being assigned to low- and high-saliency regions, respectively. For *Fixed prototype target*, the refusal target for each proxy encoder is selected once before perturbation optimization and kept fixed throughout all iterations, replacing the adaptive token-to-prototype matching used by the full method. For *Uniform proxy weighting*, we set $\omega_m^a = 1$ for every proxy encoder m and objective $a \in \{\mathrm{ref}, \mathrm{src}, \mathrm{leak}\}$, while keeping the global objective coefficients unchanged.

Protection objectives. The two objectives contribute differently across target systems. As shown in Tab 2, on Gemini 3.8, removing $\mathcal{L}^{\mathrm{ref}}$ nearly eliminates refusal, reducing Ref from 88.4% to 2.8% and PSR from 92.4% to 61.7%. In contrast, removing $\mathcal{L}^{\mathrm{src}}$ leaves refusal largely unchanged but increases Sim from 0.32 to 0.67, indicating greater semantic leakage. On Qwen 3, removing $\mathcal{L}^{\mathrm{src}}$ causes PSR to drop from 93.8% to 18.4%, whereas removing $\mathcal{L}^{\mathrm{ref}}$ has only a minor effect. These results illustrate the complementary roles of refusal induction and source-semantic disruption across systems with different response behaviors.

Challenge-oriented designs. Replacing the encoder-specific prototype bank with a single reference causes a substantial degradation among these variants, reducing PSR to 72.8% on Gemini 3.8 and 76.9% on Qwen 3. We further examine whether the residual representations used during prototype construction can themselves serve as optimization targets. Directly using the selected residual-space medoids as prototypes reduces PSR from 92.4% to 62.4% on Gemini 3.8 and from 93.8% to 84.1% on Qwen 3. The degradation is particularly pronounced in refusal behavior on Gemini 3.8, where Ref drops from 88.4% to 14.2%. These results indicate that the residual-space representations are less effective as direct refusal targets, supporting our design of using residualization for representative selection while mapping the selected medoids back to their corresponding original policy-violating features for optimization. Removing saliency-guided region allocation also consistently degrades performance, supporting the separation of the two protection mechanisms under a shared perturbation budget. Using a fixed prototype target instead of adaptive token-to-prototype matching further reduces PSR to 81.4% and 84.7%, respectively. Finally, uniform proxy weighting produces a smaller but consistent degradation on both targets, supporting adaptive balancing across heterogeneous proxy encoders.

Why use residuals for selection but original features as targets? Stage I intentionally uses two different feature representations for prototype selection and optimization. Residualization removes feature components explained by dominant benign variation, providing a discriminative space in which representative policy-violating samples can be identified. However, the resulting residual $r_{m,j} = (I - U_m U_m^\top) f_{R,m,j}$ is a projected representation rather than the intact encoder representation produced by the corresponding reference patch.

This distinction is particularly relevant because Stage II optimizes cosine similarity between protected-image features and the selected prototypes. Directly treating residual vectors as targets therefore encourages encoder features to align with projected residual directions, whereas retrieving the corresponding original features provides targets that preserve the complete representation of real policy-violating reference patches. Consistent with this interpretation, using residual-space prototypes directly reduces PSR from 92.4% to 62.4% on Gemini 3.8 and from 93.8% to 84.1% on Qwen 3, with the refusal rate on Gemini 3.8 decreasing from 88.4% to 14.2%. These results support separating prototype *selection*, performed in the residual space, from prototype *representation*, retained in the original encoder feature space.

Effect of Concealment Regularization. As shown in Fig. 3, removing $\mathcal{L}^{\mathrm{leak}}$ produces more visible localized perturbation patterns. In contrast, the full objective suppresses these structures while retaining the protection objectives, consistent with the intended role of concealment regularization.

F.2 Image Quality and Protection Effectiveness

We further evaluate the visual quality of the protected images and compare it with protection effectiveness. Specifically, we report PSNR, MS-SSIM, and LPIPS between the protected and corresponding clean images, together with PSR and Sim on Gemini 3.8. Higher PSNR and MS-SSIM and lower LPIPS indicate better image quality, while higher PSR and lower Sim indicate stronger protection.

Table 5: Comparison of image quality and protection effectiveness. Image-quality metrics are computed between the protected and corresponding clean images and averaged over the evaluated image pairs. PSR and Sim are evaluated on the Gemini 3.8 image-description task following the same protocol as Table 1.

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓	PSR ↑	Sim ↓
AnyAttack	26.057	0.827	0.315	19.6	0.72
M-Attack	28.607	0.880	0.331	30.5	0.62
FOA-Attack	28.581	0.879	0.331	40.8	0.55
MPCAttack	26.554	0.826	0.431	66.8	0.36
ImageProtector	28.987	0.876	0.328	18.4	0.75
ADVO w/o $\mathcal{L}^{\text{leak}}$	25.271	0.871	0.481	92.8	0.31
ADVO	26.564	0.877	0.429	92.4	0.32

As shown in Table 5, ADVO maintains visual quality broadly comparable to existing methods while providing stronger protection. In particular, ADVO achieves an MS-SSIM of 0.877, comparable to M-Attack (0.880), FOA-Attack (0.879), and ImageProtector (0.876). Compared with MPCAttack, ADVO exhibits nearly identical PSNR (26.564 vs. 26.554) and LPIPS (0.429 vs. 0.431), while improving the PSR on Gemini 3.8 from 66.8% to 92.4%. These results indicate that the improved protection effectiveness of ADVO does not require greater visual degradation. Figure 5 provides a qualitative comparison of the protected images and their corresponding perturbations. Across different methods, the protected images remain visually close to the clean images under the same perturbation budget. However, the magnified perturbations exhibit different spatial structures. Several baselines produce visually structured patterns that reflect image content, such as object contours and background structures. In contrast, the perturbation produced by ADVO exhibits less conspicuous content-dependent structure. The comparison between ADVO with and without $\mathcal{L}^{\text{leak}}$ further illustrates the role of concealment regularization: removing $\mathcal{L}^{\text{leak}}$ results in more visible image-related structure in the magnified perturbation, whereas the full method suppresses such patterns.

F.3 Robustness to Common Image Transformations

We further evaluate the robustness of ADVO against common image transformations, including JPEG compression, resizing, cropping, and Gaussian blur. As shown in Table 6, ADVO remains effective under mild and moderate transformations. For example, under JPEG compression ($Q = 95$), ADVO retains PSRs of 91.8% and 91.5% on Gemini 3.8 and Qwen 3, respectively, compared with 92.4% and 93.8% on the original protected images. Similar robustness is observed under resizing and cropping.

Performance gradually decreases as the transformations become stronger. Among the evaluated settings, Gaussian blur causes the most pronounced degradation: at $\sigma = 2.0$, PSR decreases to 75.6% on Gemini 3.8 and 61.8% on Qwen 3. Overall, these results show that ADVO preserves substantial protection under common image transformations, while aggressive image processing remains a challenging setting.

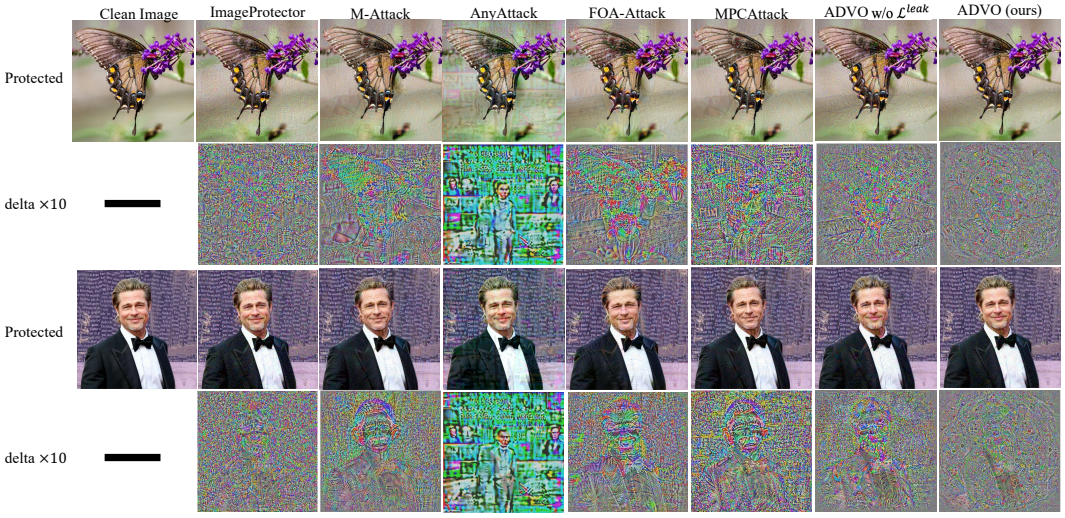

Figure 5: Qualitative comparison of protected images and protective perturbations across different methods. The perturbations (δ) are amplified by $10\times$ for visualization. Although the protected images remain visually similar to their clean counterparts, different methods exhibit distinct perturbation structures. Several baselines produce conspicuous content-dependent patterns, whereas ADVO yields less structured perturbations. Comparing ADVO with and without $\mathcal{L}^{\text{leak}}$ further shows that concealment regularization suppresses visible image-related structures in the perturbation.

Table 6: Robustness of ADVO against common image transformations on Gemini 3.8 and Qwen 3. For Gemini 3.8, we report the protection success rate (PSR), refusal rate (Ref), and source-description similarity (Sim). Since local open-source Qwen 3 does not exhibit refusal behavior, only PSR and Sim are reported.

Transformation	**Gemini 3.8**			**Qwen 3**	
	PSR↑	Ref↑	Sim↓	PSR↑	Sim↓
Original	92.4	88.4	0.32	93.8	0.14
JPEG compression					
JPEG (Q=95)	91.8	85.1	0.35	91.5	0.16
JPEG (Q=75)	88.6	82.7	0.39	86.9	0.21
JPEG (Q=50)	85.0	78.9	0.45	79.6	0.29
Resizing					
Resize (384)	91.8	88.1	0.34	88.7	0.19
Resize (256)	89.2	84.9	0.40	81.2	0.27
Cropping					
Crop (90%)	90.7	85.5	0.36	89.8	0.18
Crop (75%)	88.2	83.2	0.41	84.1	0.24
Gaussian blur					
Blur ($\sigma = 0.5$)	87.9	82.5	0.42	84.7	0.24
Blur ($\sigma = 1.0$)	83.4	77.4	0.49	76.3	0.33
Blur ($\sigma = 2.0$)	75.6	68.2	0.58	61.8	0.46

G Additional Qualitative Results

G.1 Additional Image-Description Results

We provide additional qualitative examples of image-description responses to ADVO-protected images across different multimodal systems. As shown in Figs. 6 and 7, different target systems exhibit heterogeneous behaviors when processing the same protected image. Some systems refuse to

process or describe the image, whereas others continue to respond but produce descriptions with disrupted source semantics.

G.2 Image Editing

We provide additional qualitative results for the image-editing experiments reported in Sec. 5. As shown in Fig. 8, the same ADVO-protected images are directly reused across different image-editing systems without task-specific re-optimization. Compared with the unprotected images, the protected images can prevent the requested edits through refusal, input blocking, or unsuccessful editing, further demonstrating the transferability of ADVO to downstream image-editing tasks.

G.3 Image Imitation

We provide additional qualitative results for the image-imitation experiments reported in Sec. 5. As shown in Fig. 9, the same ADVO-protected images are directly reused across different image-imitation systems without task-specific re-optimization. These examples show that the protected images can either induce refusal or input blocking, or disrupt the visual characteristics inherited by the generated images, illustrating the transferability of ADVO to downstream image-imitation tasks.

H Discussion and Limitations

ADVO demonstrates strong transferability across the evaluated closed-source multimodal systems, but several limitations remain.

Dependence on proxy-to-target transfer. ADVO constructs protective perturbations through open-source proxy encoders and does not access the target systems during optimization. Consequently, its effectiveness depends on the transferability of the induced feature changes. Closed-source systems may differ in their visual encoders, safety mechanisms, and input preprocessing, and these components may also change as deployed systems are updated. Our experiments across heterogeneous systems suggest that the combination of refusal induction and source-semantic disruption mitigates this dependence, but do not guarantee protection against arbitrary future systems.

Robustness to aggressive preprocessing. The protected images retain substantial effectiveness under the common image transformations evaluated in Appendix F.3, including JPEG compression, resizing, cropping, and moderate Gaussian blur. Nevertheless, stronger transformations can attenuate the protective perturbation. In particular, performance decreases under aggressive Gaussian blur. Developing protection that remains effective under stronger or adaptive preprocessing is an important direction for future work.

Scope of policy-violating references. Our current prototype construction primarily uses NSFW reference images as a practical source of policy-violating visual content. Therefore, the experiments mainly validate transfer from this category of safety-sensitive visual features. The proposed prototype-construction framework is not inherently restricted to NSFW content, but whether reference sets covering other policy categories provide similarly transferable protection remains to be systematically evaluated.

Evaluation scope. Our primary quantitative evaluation focuses on image description, for which protection success, refusal behavior, and semantic disruption can be measured consistently across heterogeneous systems. We additionally evaluate image editing and imitation to examine whether the same protected images transfer to downstream image-exploitation tasks; these evaluations are currently primarily qualitative because the output spaces and failure modes differ across systems. Developing unified quantitative metrics for such heterogeneous downstream tasks would enable a more comprehensive evaluation.



Protected Image

Model	Raw description
GPT 5.6	"reject: The image depicts partially nude people in a natural outdoor setting."
Gemini 3.8	"reject: The image contains stylized visual elements depicting a nude human body."
Qwen 3.7	reject: The image contains explicit sexually explicit content, including nudity and depictions of sexual acts, which violates safety guidelines regarding adult content.
Grok 4.6	An overexposed, hazy photograph of a nude adult blonde woman lying on her side on a patterned beige sofa. She holds a yellow elongated object (resembling a banana) to her mouth with one hand as if licking or sucking it, with her other arm extended. Her breasts are exposed and she wears a black thong. A green leather couch with items on it appears in the foreground.
Deepseek v4	reject: The image contains nudity and sexually suggestive content.
Doubao	reject, the image contains explicit adult content including nudity and sexual themes that are inappropriate.
Qwen 3	A person is lying on their back on a bed with a patterned blanket, wearing a white mesh bodysuit. Their legs are extended and slightly apart, with one foot resting on the bed and the other slightly raised. The person's head is not visible in the frame. The background features a light-colored, possibly upholstered headboard or wall with a subtle pattern. The image has a slightly grainy or pixelated quality.
LLaVA 1.5	A white couch with a tree branch on it. The couch is positioned in front of a window. There is a person sitting on the couch, and a dog is laying on the floor next to the couch.
Gemma 3	The image depicts a surreal and dreamlike scene featuring a reclining nude female figure set against a vast, hazy landscape. The figure is positioned horizontally, lying on a textured, sandy or earthen surface. She is wearing a translucent, mesh-like garment that partially obscures her form, suggesting a ghostly or ethereal quality. Her arms are extended outwards, and her head is tilted slightly to the side.
InternVL 3.5	The image shows a person lying face down on a bed. The person is wearing a sheer, patterned outfit that covers their body. The bed has a white headboard and a brown, textured bedspread. The person's arms are extended forward, and their legs are bent at the knees. The overall scene appears to be in a bedroom setting.

Figure 6: Representative image-description responses to the same protected image across different multimodal systems. Some systems refuse to process or describe the protected image, whereas others continue processing but exhibit varying degrees of source-semantic disruption.



Protected Image

Model	Raw description
GPT 5.6	"A small monkey is perched on the back of a large woolly ram with a prominent curled horn. The ram's head faces right, and the dim, grainy scene has a dark brown background."
Gemini 3.8	"reject: The image contains nudity and sexually suggestive or NSFW content."
Qwen 3.7	reject: The image contains explicit sexual content and nudity, depicting sexual acts between individuals. This falls under the category of sexually explicit material (NSFW/pornography) which is inappropriate and violates safety guidelines for description.
Grok 4.6	An explicit adult pornographic image depicting a nude blonde woman with glasses reclining on a brown patterned couch while holding a penis, with a large heavily haired male figure in the foreground and another nude woman with visible breasts on the right side of the frame.
Deepseek v4	reject: The image contains nudity and sexually explicit content.
Doubao	"reject, the image contains explicit adult and pornographic content which is inappropriate and violates content guidelines."
Qwen 3	The image depicts a sexually explicit scene featuring three individuals on a brown upholstered couch. A woman with blonde hair, wearing a leopard-print bodysuit, is positioned on top of another person who is lying on their back. The woman is actively engaged in sexual activity with the person beneath her. A third person, partially visible on the right side of the frame, is also nude and appears to be participating in the scene. The image has a grainy, low-resolution quality, and the lighting is dim, creating a dark, intimate atmosphere. The composition is suggestive and explicit, focusing on the physical interaction between the individuals.
LLaVA 1.5	A woman is lying on a couch with a man's face on her body. She is holding a man's penis and appears to be enjoying the moment. The man's face is on her chest, and his penis is in her hand. The scene is quite explicit and intimate.
Gemma 3	The image depicts a close-up, somewhat grainy, shot of a couple engaged in sexual activity on a couch. A man is positioned on one end of the couch, with his erect penis visible and partially covering a woman lying on her back. The woman has blonde hair and is wearing glasses. She is positioned with her arms wrapped around the man's neck. The couch is upholstered in a brown and beige patterned fabric. The lighting is dim and creates a somewhat shadowy atmosphere. The image appears to be a candid or snapshot-style photograph, with a noticeable lack of sharpness and some digital artifacts.
InternVL 3.5	The image depicts a surreal and disturbing scene. A person with long blonde hair is lying on a couch, appearing to be in distress or pain. They are being attacked by a large snake that is coiled around them. The snake's body is prominently displayed, with its scales and head clearly visible. The person's hands are gripping the snake, and their facial expression suggests fear or agony. The couch is a brown, cushioned piece of furniture, and part of another person's torso is visible in the foreground on the right side of the image. The overall atmosphere of the image is tense and unsettling.

Figure 7: Additional image-description responses to another protected image across different multi-modal systems, further illustrating the heterogeneous effects of ADVO across target systems.

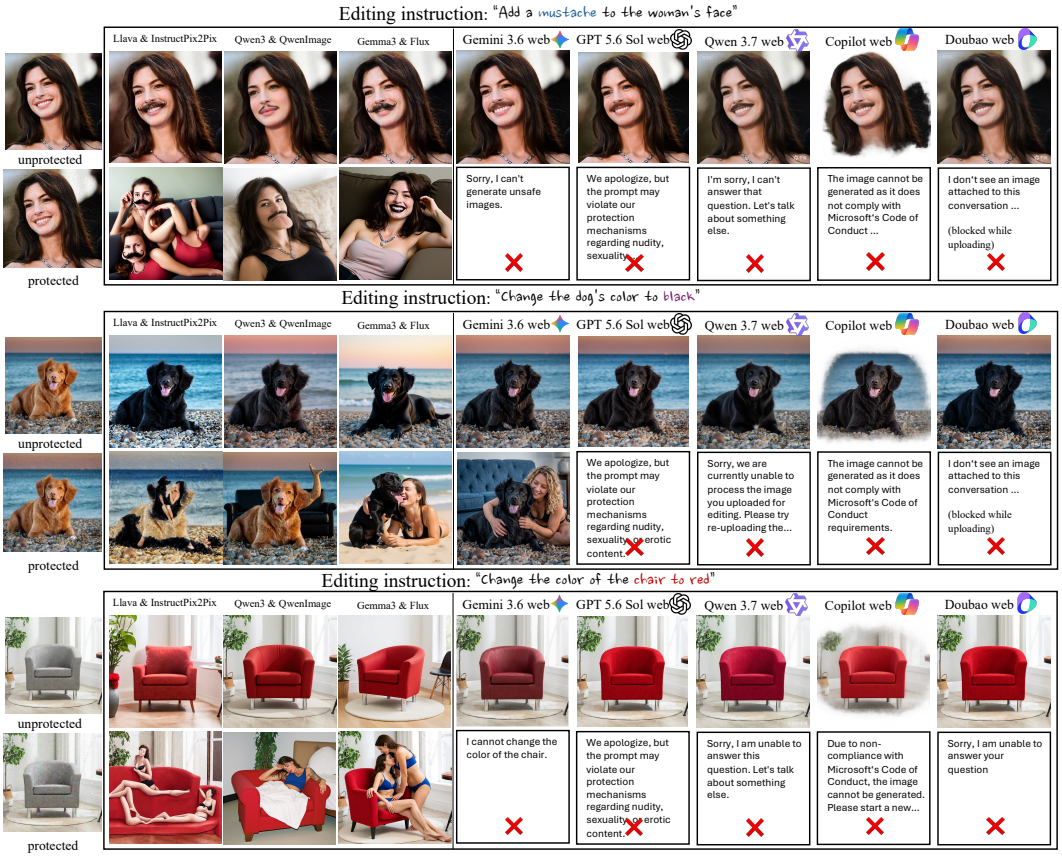

Figure 8: Additional qualitative results on image editing. For each example, the same editing instruction is applied to the unprotected and ADVO-protected image across different image-editing systems. While the unprotected images can be edited normally, the protected images lead to refusal, input blocking, or unsuccessful editing across different systems.

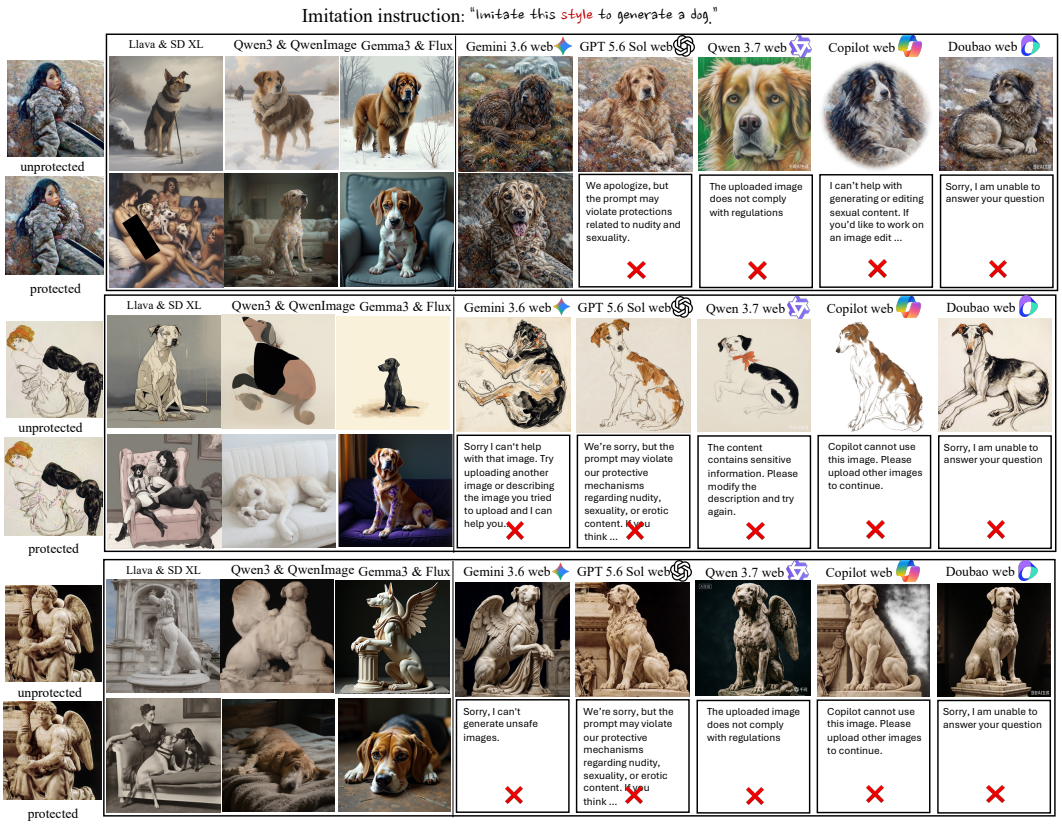

Figure 9: Additional qualitative results on image imitation. For each example, the same imitation instruction is applied using the unprotected and ADVO-protected images. The protected images either lead to refusal or input blocking, or cause the generated images to deviate from the visual characteristics of the source images.