

EditShield: Protecting Unauthorized Image Editing by Instruction-guided Diffusion Models

Ruoxi Chen¹, Haibo Jin², Yixin Liu³, Jinyin Chen^{1*}, Haohan Wang²,
and Lichao Sun³

¹ Zhejiang University of Technology, Hangzhou, China
{2112003149,chenjinyin}@zjut.edu.cn

² University of Illinois Urbana-Champaign, Champaign, USA
{haibo,haohanw}@illinois.edu

³ Lehigh University, Bethlehem, USA
{yila22,lis221}@lehigh.edu

Abstract. Text-to-image diffusion models have emerged as an evolutionary for producing creative content in image synthesis. Based on the impressive generation abilities of these models, instruction-guided diffusion models can edit images with simple instructions and input images. While they empower users to obtain their desired edited images with ease, they have raised concerns about unauthorized image manipulation. Prior research has delved into the unauthorized use of personalized diffusion models; however, this problem of instruction-guided diffusion models remains largely unexplored. In this paper, we first propose a protection method **EditShield** against unauthorized modifications from such models. Specifically, **EditShield** works by adding imperceptible perturbations that can shift the latent representation used in the diffusion process, tricking models into generating unrealistic images with mismatched subjects. Our extensive experiments demonstrate **EditShield**'s effectiveness among synthetic and real-world datasets. Besides, we found that **EditShield** performs robustly against various manipulation settings across editing types and synonymous instruction phrases.

Keywords: Diffusion models · Text-to-image · Image editing

1 Introduction

In the domain of image generation, the emergence of diffusion models [20, 35, 39] has marked a new era of innovation, especially in text-to-image generation tasks. These large-scale models stand out for their ability to produce photorealistic and artistic images. Notably, Stable Diffusion (SD) [33] and DALL-E 3 [4] are capable of generating compelling images from simple text prompts in seconds. Their widespread adoption fuels millions of applications in areas ranging from digital art to content creation.

* Corresponding author

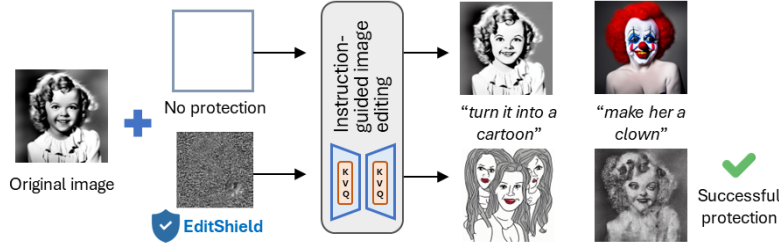


Fig. 1: The illustration of protection by **EditShield**. Editors cannot get their expected images with the protection of **EditShield**.

These text-to-image diffusion models have also emerged as powerful tools for image editing [3, 23, 38, 43]. Instruction-driven image editing models [7, 15, 21, 46] represent a significant advancement, can modify the given source image according to the instructions specified by the user. Remarkably, these advancements facilitate modifications directly during the forward pass, eliminating the need for additional inputs such as reference images, segmentation masks, or model fine-tuning. This flexibility also benefits practicality as such guidance is more aligned with human intuition.

The success of the instruction-guided image editing also presents avenues for misuse with serious ethical implications. The malicious editor may create misleading narratives or “fake news”, thereby contributing to public misinformation [14, 41]. Similarly, unauthorized modifications of celebrity images or personal photographs—ranging from altering facial expressions to creating inappropriate content - pose grave concerns for privacy, consent, and reputational harm. Artists may see their works manipulated without permission [22], leading to misunderstandings, economic losses, and a diminution of artistic integrity [12, 40].

Several attempts [25, 36, 37, 42] have been proposed to address the unauthorized modifications by diffusion models and have shown effectiveness through experiments. These solutions, however, focus on preventing personalized editing, which requires reference images and the fine-tuning process. The protection benefits from fixed and pre-defined text prompts used in personalization, facilitating the development of specific countermeasures. In contrast, instruction-guided image editing, characterized by its dynamic and user-specific instructions, presents a significant challenge to existing protective approaches. This distinction underscores the urgent need for innovative research focused on developing robust protections for instruction-guided editing, regardless of editing types and instruction phrases.

To tackle these challenges, in this paper, we investigate how to protect images against unauthorized editing from instruction-guided diffusion models, as depicted in Fig. 1. We design **EditShield**, a protection method designed to safeguard images from being unauthorized edited by instruction-guided diffusion models. It operates by introducing strategic perturbations that effectively disrupt the latent representations these models rely on, leading to significant

mismatches between the intended editing outcomes and the actual results, often manifesting as compromised image quality and mismatched subjects. Through extensive experiments on both synthetic and real-world datasets, we verify the effectiveness of **EditShield** across diverse images and is robust against various manipulation settings under different editing types and synonymous instruction.

Our contributions are summarized as follows:

- We take the first step to study the potential risk of unauthorized image editing from instruction-guided diffusion models. We then develop **EditShield**, to safeguard images from unauthorized modifications by these models.
- EditShield works by introducing small perturbations that can disrupt the latent representation leveraged by those models, resulting in unrealistic images with mismatched subjects after editing.
- Through extensive experiments, we show that our method effectively protects a wide range of images from unauthorized editing and exhibits robustness against diverse editing types and instruction phrases.

2 Background

2.1 Diffusion Models

Motivated by the non-equilibrium statistical physics, diffusion models [18, 20, 27, 32, 35, 39] have shown superior performance in text-to-image generation tasks, notably excelling in producing images of remarkable quality and diversity. Typically, they employ a pre-trained model such as CLIP [30] or T5 [31] for extracting text embeddings, which are then leveraged by the generation component for synthesizing images.

These models distinguish themselves from language models by utilizing continuous latent representations grounded in the Markov Chain framework. Given an input image $x_0 \sim q(x)$, the forward diffusion process adds a small amount of Gaussian noise to the sample in T step that produces a sequence of noisy samples $\{x_1, x_2, \dots, x_T\}$. The step sizes are controlled by a variance schedule $\{\beta_t \in (0, 1)\}_{t=1}^T$. We can sample x_t at any arbitrary time step t , as follows:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon \quad (1)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Conversely, the backward process uses the reverse Markov Chain to denoise the sample x_t from x_{t+1} . The injected noise in each step can be learned by the denoising model ϵ_θ , which is trained to minimize the distance between the estimated noise and the true noise.

With the integration of a Variational Auto-Encoder (VAE) [24] trained on images, latent diffusion models (LDMs) [29, 33] significantly enhance the efficiency and quality of the generated images while concurrently reducing computational costs. They are now widely employed in visual generation [2, 5, 6] and editing [10]. Given an image x , the image encoder $\mathcal{E}$ first encodes it into a latent representation $z = \mathcal{E}(x)$. Then, the noise-adding and denoising operations are conducted on encoded latent z . The unconditioned latent diffusion models are trained:

$$\mathcal{L}_{unc} = \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0, \mathbf{I}), t} [\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2] \quad (2)$$

where $\mathcal{L}_{unc}$ denotes the loss function of unconditional diffusion models, t is the time step, and z_t is the latent noise at time step t .

2.2 Instruction-guided Image Editing

Text-guided image editing [3, 9, 11, 23, 38, 43] have demonstrated the efficacy of leveraging pre-trained diffusion models, in the nuanced realm of image editing such as subject-specific modifications and style transfers. Besides, some models [16, 34] enable users to achieve highly customized edits, through fine-tuning with some reference images and text prompts.

Different from them, instruction-guided image editing accepts straight commands like “add fireworks to the sky” for guidance, without the need for additional descriptions or regional masks. Brooks et al. [7] first proposed Instruct-Pix2Pix based on Prompt-to-Prompt [19]. The model leverages texts generated by a fine-tuned version of GPT-3 [8] and images generated by SD. Subsequent works are all built upon InstructPix2Pix and further enhanced the performance by fine-tuning it with the MagicBrush dataset [46], incorporating human feedback [47] for better alignment, and with the assistance from multi-modal large language models [15, 21]. These models are based on latent diffusion and introduce input images and instruction prompts for guidance. Given the image conditioning c_I and text conditioning c_T , the network is trained to predict the noise added on the latent, for image content preserving and instruction following. The objective of latent diffusion is minimized as follows:

$$\mathcal{L}_{ins} = \mathbb{E}_{\mathcal{E}(x), \mathcal{E}(c_I), c_T, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T)\|_2^2] \quad (3)$$

where $\mathcal{L}_{ins}$ denotes the loss function of instructional image editing models.

As Fig. 2 suggests, the user inputs a source image and an instruction prompt to the model, where the image and text are encoded by the image and text encoder or MLLMs, respectively. The latent diffusion model is trained under the guidance of both image embedding and text embedding. Finally, the edited image will be reconstructed by the image decoder from its latent representation.

In this paper, we mainly target InstructPix2Pix [7] model, and assess the generalization of our method to its enhanced fine-tuned version on MagicBrush [46], since these models are open-sourced and publicly available until now.

2.3 Protection Against Unauthorized Use

As diffusion models rapidly evolve, concerns regarding its misuse cannot be overlooked. To address this issue, various strategies have been developed to safeguard public images of individuals against diffusion models. The main idea involves introducing specially crafted noise to the image. For instance, PhotoGuard [36] introduces imperceptible perturbations to images, effectively raising the computational cost for malicious editing attempts. As a result, LDMs cannot generate realistic images when attempting to edit upon the perturbed image. In the field of artworks and paintings, Shan et al. [37] developed Glaze which can protect

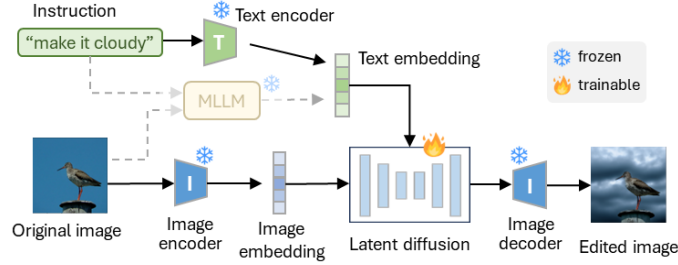


Fig. 2: The workflow of instructional image editing based on diffusion models.

artists against style mimicry by text-to-image diffusion models. Similarly, Liang et al. [25] adopted the concept of adversarial examples for protecting images from being learned, imitated, and copied by diffusion models. Van Le et al. [42] generated subtle noise to the image published by the user, to disrupt the generation quality of DreamBooth trained on these perturbed images. These protective measures target copyright infringers who utilize diffusion models with specific prompts tied to particular subjects. However, their effectiveness is challenged in real-world scenarios where models encounter a variety of unseen prompts, potentially making these protections ineffective. Furthermore, the specific problem posed by instruction-guided diffusion models—where user-generated instructions dictate the content creation process - remains unexplored by existing works.

3 Preliminary

3.1 Problem Statement

We first introduce the detailed protection setting and assumptions, where we consider a two-player game: the Image Editor for instructing diffusion model and the Image Protector for the protection method.

Protection Scenario. Instruction-guided editing models can be convenient for modifying images. But they may also be exploited by unauthorized use towards a specific individual or community. Within this context, we identify three key entities in the lifecycle of information generation, modification, and consumption: the photo owner, who creates or is depicted in the images; the malicious editor, who intends to modify the image for unauthorized use for malicious purposes, such as spreading fake news; and the consumer, who may unknowingly encounter and believe the manipulated content. This flow from content creation to its unauthorized modification and eventual consumption is depicted in Fig. 3. To counteract these risks, the protector helps shield images from such manipulations, thereby making unauthorized edits ineffective.

Image Editor. The editor’s goal is to modify the image from the image owners without permission. The image editor can access the well-trained, instruction-guided image editing models and then exploit them to modify images for po-



Fig. 3: The image editing scenario that we consider.

tentially harmful purposes, such as negatively impacting one’s appearance. Furthermore, the editor can obtain images from the social platform from their own-ers, bypassing any permissions or ethical considerations. The editing process is straightforward: the editor feeds the source image into the model and provides specific instruction prompts.

Image Protector. The goal is to add noise to the image to prevent unauthorized editing. The protector can use our method to generate perturbations and inject them into the image before publishing it online. The protection is successful if the image has a mismatched subject identity or is of awful quality after editing. The protector can be the image owner or another third-party service.

We assume the protector is aware that the editor will use instruction-guided diffusion models for image editing. The protector has white-box access to those editing models. This assumption is now common among works [25, 36] since more models are open-sourced online. However, the protector does not know the instructions and editing type specified by the editor. Besides, we assume the protector has significant computational power for generating perturbations.

3.2 Design Challenges

Existing solutions [25, 42], have showcased effectiveness in scenarios involving personalized image editing [16, 34], where protections are designed around specific prompts tied to given reference images and pre-defined subjects. These personalized image editing models operate under the premise of fixed text prompts, like “a photo of sks [class noun]”, where sks specifies the subject and “[class noun]” is the subject category. However, these protection mechanisms face challenges in instruction-guided editing scenarios, where the image protector lacks access to the dynamic instruction prompts and editing types employed by unauthorized editors. The absence of specific prompts poses a substantial challenge in designing an effective protection strategy.

4 Our Proposed Method

4.1 Design Intuition

With unknown instructions that unauthorized editors might use, designing protection perturbations may be difficult. Given that the input images are accessible to the protector, our solution explores the potential within the image space itself.

Image editing aims to enact desired modifications through language commands while retaining as much of the original image’s detail as possible. The unauthorized editing aims to modify images without losing inherent content and quality. For example, an instruction to “*replace the dog with a cat*” should not accidentally transform the surrounding lawn into a desert. Such unintended changes, while not fulfilling the editor’s intent, present an opportunity for protection by exploiting these mismatches.

Inspired by this, we propose leveraging content changes as a protection mechanism against malicious editing, independent of the specific instructions. Our approach involves crafting protective noise designed to disturb the editing process guided by a spectrum of potential instructions.

We target the latent representation that is leveraged by the mainstream of instruction-guided diffusion models. We introduce a instruction-independent protection method **EditShield** aims to shift the latent representation of the original image while maintaining the protection noise imperceptible. After that, following the instruction-guided diffusion process, the edited image will be irrelevant or unrealistic, effectively safeguarding against unauthorized modifications.

4.2 EditShield

Based on our previous analysis, we derive the two conditions that the protected image should satisfy, then formulate **EditShield**, as an optimization problem based on those two conditions.

Latent inconsistency. Recall that in instruction-guided image editing models, image conditioning and text conditioning are determined by the source image and the instruction prompt, respectively. So we can manipulate the source image to influence the image embedding. This can be achieved by shifting the latent representation before entering the latent diffusion process. In this way, the protection operates independently of the instruction prompt. Even if the editing follows the instructions provided by an unauthorized editor, it is not possible to obtain an edited image that maintains related subjects.

Note that the model first projects the source image into the image representation by VAE. To disrupt the image embedding in the latent space, we force the projected embedding of the protected image to deviate from the original representation. We define the latent inconsistency loss to quantitatively measure the latent shift: $\mathcal{L}_{latent} = Dist((\mathcal{E}(x_p), \mathcal{E}(x)))$, where x_p and x represent the protected and the source image, respectively. $\mathcal{E}$ is the VAE used in the instruction-guided editing model. $Dist(\cdot, \cdot)$ computes the distance between two latent representations. We use mean-square error (MSE) by default.

Perceptual consistency. While shifting latent representation, the protective noise should not hurt the visual quality of the source image. So, the protection should be human-invisible and should not affect image semantics. Visible noise could alert an unauthorized editor to the presence of protection measures, prompting them to either discard the image, conduct data transformations such as filtering, or apply specific high-resolution editing instructions to circumvent

the protection. Therefore, the noise must be well-crafted to be invisible to humans, ensuring that the image retains its original appeal and message while still effectively disrupting unauthorized editing processes. We define the perceptual consistency loss as $\mathcal{L}_{perceptual} = \|x_p - x\|_2^2$ with l_2 norm regularization.

Transformation Robustness. The shift in latent representation, while effective in preventing unauthorized image modifications, can introduce mismatches in content and unrealistic textures, compromising the edited image’s integrity. However, perturbations crafted directly with the above two conditions are fragile to minor transformations. To address this, we incorporate the Expectation Over Transformation (EOT) [1], into our optimization strategy. Specifically, given F as a distribution over a set of transformations f , we integrate the transformation robustness loss into $\mathcal{L}_{latent}$ and rewrite it as $\mathcal{L}_{latent} = \text{Dist}((\mathcal{E}(f(x_p))), \mathcal{E}(x))$, where $f(x_p)$ is the transformed image of x_p after applying the transformation f . This integration ensures that our protective perturbations remain effective and consistent across a variety of potential editing scenarios, including those involving subtle image transformations.

Combining the above losses together, our final optimization problem is:

$$\max_{x_p} \mathbb{E}_{f \sim F} [\text{Dist}((\mathcal{E}(f(x_p))), \mathcal{E}(x))] - \beta \cdot \|x_p - x\|_2^2 \quad (4)$$

where β is the hyper-parameter to balance the two losses. To guarantee that the perturbation is human-invisible, the protection noise δ is bounded by the overall perturbation budget. We use projected gradient descent to solve the optimization problem efficiently. In practice, we initialize x_p by x plus an extra small standard Gaussian noise $\mathcal{N}(0, \mathbf{I})$. This iterative process persists until the protected image reaches convergence or when the maximum number of iterations S is obtained. After applying the perturbation generated by **EditShield**, users can publish their images online. It can be seen from the above that **EditShield** generates untargeted perturbations. We don’t set a specific target before optimization to make it more general towards input images consisting of different contents.

5 Evaluation

5.1 Experiment Setup

Models. Our objective is to prevent image editing from being used by instruction-guided diffusion models without permission. We choose *InstructPix2Pix* (*ip2p*) [7], which is the mainstream tool and basic backbone for instruction-guided image editing. We use the *ip2p* model based on SD v1.5, along with its enhanced model *ip2p-mb* that is fine-tuned on MagicBrush dataset [46].

Datasets. We evaluate the effectiveness of our protection using two datasets: a filtered subset from the synthetic dataset IPr2Pr [7] and the test split of MagicBrush. The protection is applied to the source images, and instructions provided in the dataset are used for subsequent image editing.

Baselines. We compare our method with PhotoGuard [36], which adds imperceptible perturbations to prevent image editing with text prompts. The parameters are configured following the settings reported in the original paper.

Evaluation Metrics. Following [7], we calculate two metrics to assess image editing, as discussed below.

- CLIP image similarity: We first extract image embedding from both the source and edited images without and with protection. We then calculate the cosine similarity of these CLIP image embeddings. This metric measures how much the edited image agrees with the source image.
- CLIP text-image direction similarity [17]: We adopt LLaVA-1.5 13B [26] for generating descriptions of both the source and edited images without and with protection. These descriptions, combined with the editing instructions, are used to compute CLIP text-image direction similarity. This metric measures how much the change in text captions agrees with the change in the images.

For both metrics, smaller values indicate worse image editing capabilities and, consequently, better image protection capability.

Implementation Details. We apply protections on 2,000 source images from IPr2Pr and 1,000 images from MagicBrush. Subsequently, these images are edited by instruction-guided diffusion models, with the guidance of the corresponding instructions provided in each dataset. For T in EOT, we choose transformations that the editor might use, including Gaussian kernel smoothing (kernel=5, sigma=1.5), image rotation with rotate angle=5°, and center cropping. We set the overall perturbation budget as $4/255$, $\beta = 0.2$, and $S=30$. To mitigate randomness, we use the same seed when editing source and protected images. We defer further experimental details to supplementary materials.

5.2 Qualitative Results on Protection

We provide a qualitative analysis of various images in Fig. 4. The source images and their corresponding instructions in the uppermost four rows are from IPr2Pr, while the last two are from MagicBrush. We compare the quality of editing on images without the protection, and with the protection of **EditShield**. It can be observed that images derived from protected images exhibit mismatched subjects and low quality compared with those edited from the original source images. For instance, in the first row, under the instruction “*turn into a cartoon*”, the girl in the protected image disappears, replaced by a group of cartoon characters. In the last row, the background of the edited image is heavily distorted with blue noise. These results demonstrate that an unauthorized editor is unable to achieve the desired output, thus underscoring the effectiveness of our protection.

5.3 Quantitative Results on Protection

Protection Effectiveness. We calculate two metrics to assess image protection: the cosine similarity of CLIP image embeddings and the directional CLIP similarity, both applied to unprotected and protected images by PhotoGuard and **EditShield** under the same instruction. Results are shown in box plots in Fig. 5, where orange, green and blue boxes represent similarities without protection, with PhotoGuard and **EditShield**, respectively. The box shows the quar-

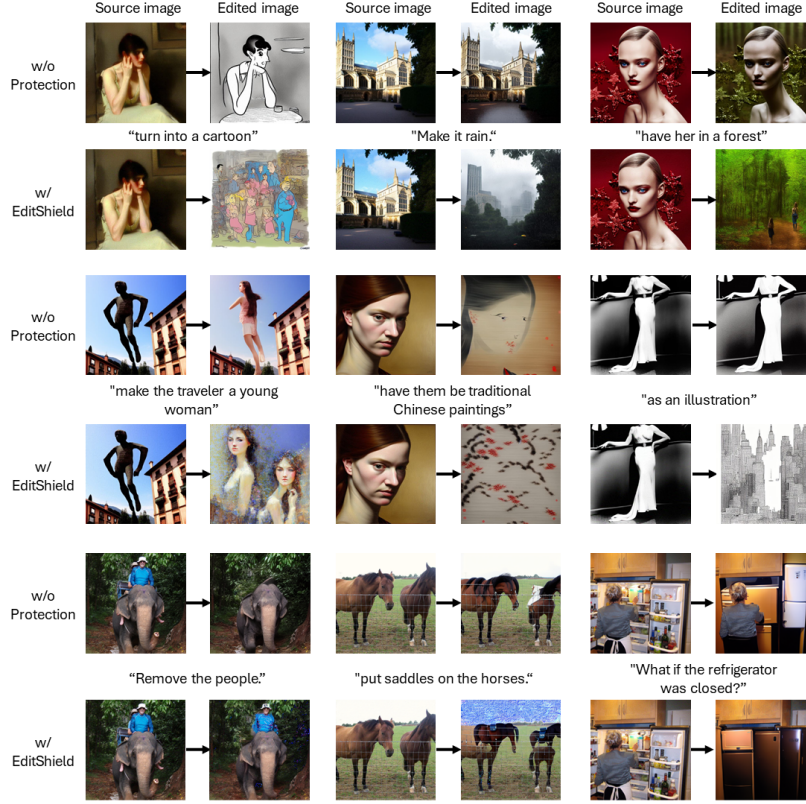


Fig. 4: Qualitative protection results. Each pair of rows displays the source and protected images with their respective edited versions.

tiles of the data while the whiskers extend to show the rest of the distribution, except for points that are “outliers”.

As indicated by Fig. 5, **EditShield** performs better than PhotoGuard, as evidenced by the smaller values in the blue boxes across both similarity metrics. When applying **EditShield**, the image editing is less successful. In other words, changes made during the editing of the protected image show less agreement with the given instruction. Furthermore, the protected image after editing exhibits more differences in image content compared with those without protection. This is because **EditShield** can shift the latent representation of the original image, resulting in lower image consistency after editing.

Image Quality Assessment. To evaluate image quality, we compare the editing results without protection, and with protection by PhotoGuard and **EditShield**, including Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [44]. For both metrics, lower scores indicate worse image quality after editing. Results are presented in Table 1. As observed, with the protection

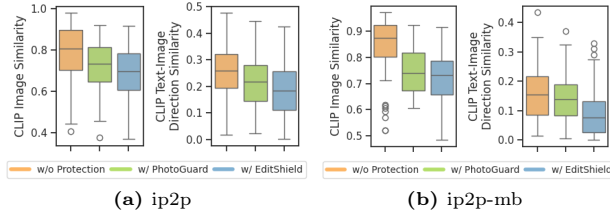


Fig. 5: Quantitative results on editing with protection of PhotoGuard and EditShield.

Model	Protection	PSNR	SSIM
ip2p	w/o Protection	4.11	0.14
	w/ Photoguard	4.06	0.11
	w/ EditShield	3.32	0.09
ip2p-mb	w/o Protection	19.11	0.78
	w/ Photoguard	17.56	0.67
	w/ EditShield	16.82	0.66

Table 1: Image quality comparison.

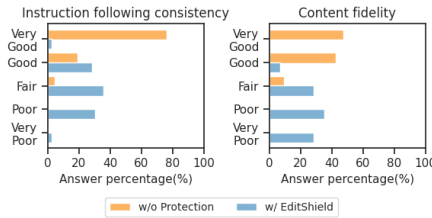


Fig. 6: Editing evaluation by GPT-4V.

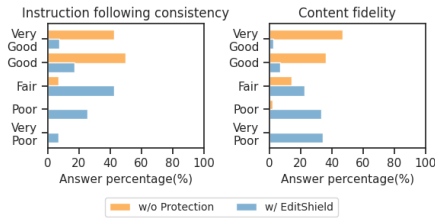


Fig. 7: Editing evaluation by humans.

of **EditShield**, edited images consistently show smaller quality scores. Compared with PhotoGuard, **EditShield** can result in poorer image quality.

Evaluation by GPT-4V. We further leverage GPT-4V [28] to comprehensively evaluate the instruction-guided image editing with the protection of **EditShield**. It is asked to give quantitative scores on instruction-following consistency and content fidelity. The former represents how much the edited image faithfully follows the given instruction, while the latter evaluates the consistency in subject, style, content, and image quality between the source and the edited image. Both scores are measured on a scale from 0 to 1. For a more specific analysis, we further categorize these scores using a 5-level Likert scale, representing “very poor”, “poor”, “fair”, “good” and “very good”.

We randomly choose 100 “source-edited” image pairs with **EditShield** and report evaluation results in Fig. 6. As observed, when **EditShield** is applied, GPT-4V’s scores drop to a great extent, with a notable increase in scores categorized as “very poor” and “poor” based on content fidelity and instruction-following consistency. These findings align with our quantitative assessment, underscoring **EditShield**’s success in preventing unauthorized image modifications by compromising the consistency and fidelity of the edited image. As a result, images edited with the protection of **EditShield** exhibit mismatched subjects or unrealistic contents, effectively preventing unauthorized editors from achieving their intended modifications. We will provide details of querying GPT-4V in the Supplementary materials.

Evaluation by Human Users. We also conduct a human evaluation for comparative analysis. Twenty randomly selected “source-edited” image pairs, along

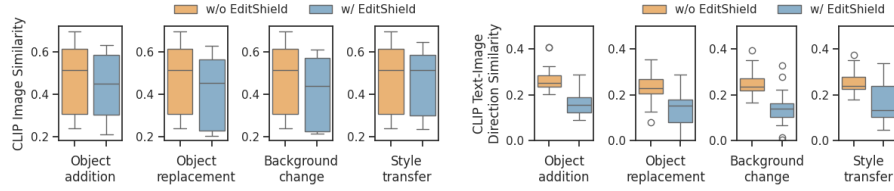


Fig. 8: The impact of different types of instructions on EditShield.

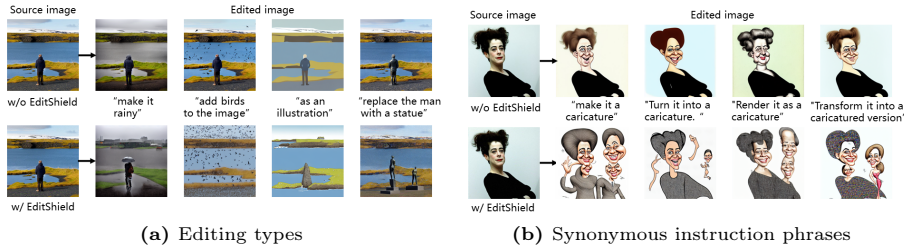


Fig. 9: Edited images with EditShield on different instructions.

with corresponding instructions, are presented to 50 participants. Each participant is tasked with assessing whether the edited image meets the editing requirements, considering content fidelity, instruction following, and image quality. Quantitative scores between 0 and 1 will be given by those participants. We calculate the percentage of five responses and report results in Fig. 7. We observe that most images, with the protection of **EditShield**, don't align with the editing requirements. In other words, **EditShield** effectively prevents editors from achieving their intended editing goals. This outcome not only corroborates the quantitative efficacy of our protection as previously demonstrated but also highlights its practical effectiveness in preserving images against unauthorized modifications. Further details are in Supplementary materials.

5.4 Protection Robustness

We verify the protection robustness of **EditShield** under different instruction prompts, including different editing types and synonymous instruction phrases.

Editing Types. We further investigate how the editing instruction types will affect **EditShield**. According to [46], we focus on four types of instructions: object addition, object replacement, background change and

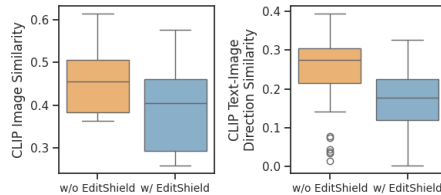


Fig. 10: The impact of synonymous instruction phrases on EditShield.

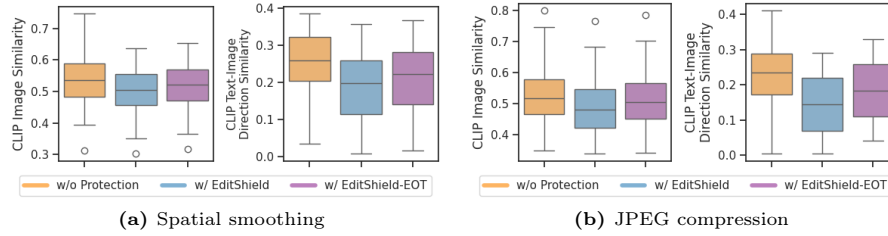


Fig. 11: Quantitative results on possible countermeasures.

style transfer. We randomly chose 100 images and conducted different types of image editing on them. Two similarity metrics are reported in Fig. 8. Besides, we also present some visual analysis in Fig. 9a. We can observe **EditShield** shows robust protection performance against a variety of instruction types, as evidenced by all lower values in blue boxes across both similarity metrics. Notably, the protection is more effective for instructions involving object changes, such as object addition and replacement. This can be attributed to **EditShield**’s ability to induce mismatches in subjects. **Synonymous Instruction Phrases.** We consider how **EditShield**’s performance might vary with synonymous instruction phrases, such as using “*make it burn*” instead of “*make it on fire*”. We leverage ChatGPT to generate four synonymous phrases for each original instruction prompt. Then we perform image editing on protected images for comparison. For measurement, we calculate image and direction similarity on average, then present quantitative results and visual examples in Fig. 9 and Fig. 10, respectively. We observe that the values in the blue boxes remain lower than those in the orange boxes. This suggests that using synonyms in instruction expressions will not affect the effectiveness of **EditShield**’s protection.

5.5 Possible Countermeasures

We consider potential countermeasures the unauthorized editor might employ to reduce our protection effectiveness.

One potential approach involves leveraging image processing methods for protection mitigation, which does not require retraining the generative model or modifying the model structure. After applying **EditShield**, spatial smoothing (SS) [45] and JPEG compression (JC) [13] may be adopted by editors to mitigate the protection. Here we set window *size* = 2 in SS and *quality* = 80 in JC. We calculate two similarity metrics for editing evaluation. Besides, we also remove EOT in **EditShield** for comparison. Results are reported in Fig. 11. As observed, two metrics drop after our protection, i.e., **EditShield** still shows effectiveness in image protection when facing these image transformations. When we remove EOT, protection effectiveness decreases, as purple boxes exhibit larger metrics

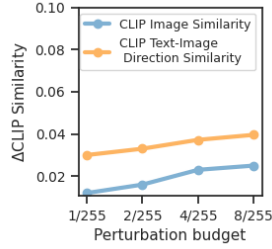


Fig. 12: Impact of overall perturbation budget.

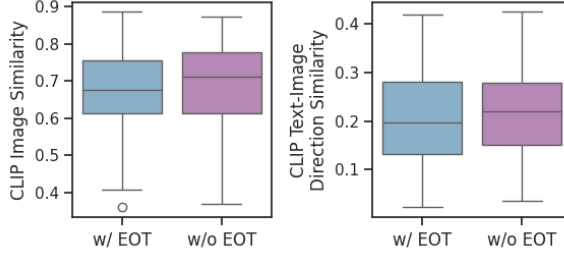


Fig. 13: Impact of EOT.

than blue ones. Such observations demonstrate that the integration of EOT can improve **EditShield**’s protection against image transformations.

5.6 Ablation Studies

Perturbation Budget. We investigate the effect of the overall perturbation budget on our protection. Utilizing the IPr2Pr dataset, we explore how varying budgets from 1/255 to 8/255 influence the protection’s effectiveness, as measured by the median drop in two critical similarity metrics. Results are shown in Fig. 12. We observe that a relatively modest perturbation budget of 4/255 is sufficient for **EditShield** to significantly prevent the editing performance of instruction-guided image editing models, thereby leading to robust protection against unauthorized modifications.

Impact of EOT. To evaluate the individual contributions of EOT to its overall effectiveness, we conducted ablation studies using the IPr2Pr dataset. The results, visualized in Fig. 13, show blue boxes are slightly lower than purple ones in two metrics. This indicates that the integration of subtle image transformations in **EditShield** contributes to more effective protection.

6 Conclusion

In this paper, we propose the first protection method against unauthorized editing by instruction-guided diffusion models. The proposed protection **EditShield**, crafts small perturbations that try to shift the latent distribution from the source image. When the protection is applied, the edited image will have mismatched subjects and be of low quality. Extensive evaluations of two datasets have verified that **EditShield** can effectively prevent images from being edited by instruction-guided diffusion models. Besides, **EditShield** is robust towards different editing types and synonymous instruction phrases. However, **EditShield** shows limitations in protecting images in a black-box manner against online instruction-guided image editing platforms. An important future direction is to design a more efficient and robust perturbation in practice.

Acknowledgements

We extend our gratitude to all authors, reviewers, and the chair for their invaluable contributions. Additionally, we would like to express our appreciation to Dongping Chen for providing computational resources.

References

1. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: International conference on machine learning. pp. 284–293. PMLR (2018)
2. Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)
3. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18208–18218 (2022)
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. <https://cdn.openai.com/papers/dall-e-3.pdf> **2**(3), 8 (2023)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
9. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Advances in Neural Information Processing Systems **36** (2024)
10. Corneanu, C., Gadde, R., Martinez, A.M.: Latentpaint: Image inpainting in latent space with diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4334–4343 (2024)
11. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
12. Dean, I.: The raw impact of ai on artists’ lives. <https://www.creativebloq.com/features/ai-art-the-impact-of-generative-AI> (2023)
13. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpeg compression on adversarial images. arXiv preprint arXiv:1608.00853 pp. 1–8 (2016)
14. Felix M. Simon, Sacha Altay, H.M.: Misinformation reloaded? fears about the impact of generative ai on misinformation are overblown. <https://misinfoforeview.hks.harvard.edu/article/misinformation-reloaded-fears-about-the-impact-of-generative-ai-on-misinformation-are-overblown/> (2023)

15. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. arXiv preprint arXiv:2309.17102 (2023)
16. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023)
17. Gal, R., Patashnik, O., Maron, H., Bermano, A.H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* **41**(4), 1–13 (2022)
18. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
19. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
21. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., et al.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. arXiv preprint arXiv:2312.06739 (2023)
22. Jiang, K.: These ai images look just like me. what does that mean for the future of deepfakes? *Toronto Star*
23. Kavar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
25. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. Proceedings of Machine Learning Research*, vol. 202, pp. 20763–20786. PMLR (2023)
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
27. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: *International Conference on Machine Learning*. pp. 16784–16804. PMLR (2022)
28. OpenAI: Openai models - gpt-4-vision. <https://openai.com/research/gpt-4v-system-card> (2023)
29. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

- natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**(1), 5485–5551 (2020)
 32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* **1**(2), 3 (2022)
 33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10684–10695 (2022)
 34. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22500–22510 (2023)
 35. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022)
 36. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588* (2023)
 37. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models. In: Calandrino, J.A., Troncoso, C. (eds.) *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*. USENIX Association (2023)
 38. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089* (2023)
 39. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: *International conference on machine learning*. pp. 2256–2265. PMLR (2015)
 40. Tolcheva, S.: The dark side of ai art: 4 potential issues with the growing trend. MUO
 41. TUCKER, J.A.: Ai could create a disinformation nightmare in the 2024 election. <https://thehill.com/opinion/4096006-ai-could-create-a-disinformation-nightmare-in-the-2024-election/> (2023)
 42. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N., Tran, A.: Antidreambooth: Protecting users from personalized text-to-image synthesis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2116–2127 (2023)
 43. Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., et al.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18359–18369 (2023)
 44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
 45. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: *25th Annual Network and Distributed System Security*

- Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018. pp. 1–15. The Internet Society (2018)
46. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. arXiv preprint arXiv:2306.10012 (2023)
47. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: Hive: Harnessing human feedback for instructional visual editing. arXiv preprint arXiv:2303.09618 (2023)