EIRES:Training-free AI-Generated Image Detection via Edit-Induced Reconstruction Error Shift

Supplementary Material

1. Summary

For clarity and completeness, we provide additional details of our method in the supplementary material. The supplementary document is organized as follows:

- In Section 2, (i) we provide the pseudocode and an illustrative example of the EIRES detection pipeline, and (ii) additional details of the experimental setup.
- In Section 3, we show visual comparisons between recent raw reconstruction error methods and our proposed EIRES, highlighting error distribution histograms on the GenImage dataset and under JPEG compression and cropping.
- In Section 4, we provide the complete proofs of **Proposition 1** and **Proposition 2** from Section 4.3 of the main paper, covering the lower bound of raw reconstruction error and the lower bound under edit perturbation.
- In Section 5, we discuss the limitation of our method and outline our future research.

2. Method and Implementation Details

2.1. EIRES Detection Pipeline

The EIRES detection pipeline is summarized in the pseudocode provided in Algorithm 1. Given an input image, either real or generated, the method applies a predefined set of structured edits and performs autoencoder-based reconstruction for each edited version. The reconstruction deviations relative to the original image are then computed, and the maximum deviation across all edits is used as the final detection score. To complement the pseudocode, Figure 1 offers a visual example of the detection workflow, illustrating how structured edits are generated and how the corresponding reconstruction shifts are measured.

Algorithm 1 EIRES Detection Pipeline

Input: input image or a set of images x
Output: detection result `Label`
Notation:

- $\mathcal{S}$: the set of structured editing operations
- $T(\cdot)$: structured editing operation
- $f(\cdot)$: pre-trained AutoEncoder
- $d(\cdot,\cdot)$: LPIPS perceptual distance metric

$\tilde{x} \leftarrow T(x), \quad T \in \mathcal{S}$
$e_{\text{pre}} \leftarrow d(x, f(x))$
$e_{\text{post}} \leftarrow d(\tilde{x}, f(\tilde{x}))$
$\triangle e^* \leftarrow \max_{T\in\mathcal{S}}\{e_{\text{post}} - e_{\text{pre}}\}$
if $\triangle e^* < \tau$ **then**
  `Label` $\leftarrow$ real
else
  `Label` $\leftarrow$ fake
end if
return `Label`

2.2. Implementation Details

In our main experiments and ablation studies, we employ the ERNIE-iRAG-Edit model from the Baidu Qianfan platform [3] and report results under the Multi-Edited strategy. For Add, we apply a circular mask of 75-pixel radius at the image center with the prompt "Add an object in this area". The Erase operation uses the same mask with the prompt "Erase the masked area and fill naturally". For SemR, no mask is used and the prompt is "Reconstruct as original". For experiments in Section 5.3 (Results on Unbiased Dataset) and Section 5.5 (Robustness to Unseen Perturbations), which require evaluating performance under multiple crops and JPEG compression levels. To reduce API cost at scale, we switch to the open-source FLUX.1 model [4] and apply the simplest Sem editing strategy using the same prompt. As shown in Figure 2, FLUX.1 produces stable and controllable edits comparable to those of commercial editing engines, ensuring reliable evaluation in large-scale experiments.

The implementation code is provided in the supplementary material. All experiments are conducted on a workstation running Ubuntu 20.04.6 LTS with Python 3.10.16 and an NVIDIA RTX A6000 GPU. We implement EIRES using PyTorch 2.1.2 with CUDA 12.1, and employ NumPy 1.26.3 for numerical computation.

Although our experiments involve applying edits to a large number of images, which leads to non-negligible API and computational cost, this overhead comes entirely from the requirements of experimental evaluation rather than from the EIRES method itself. In real-world deployment, EIRES only needs to apply editing operations to the images being checked, and such edits can be generated efficiently. Modern commercial editing engines (e.g., Doubao and Midjourney [1, 2]) can produce high-quality structured edits rapidly and at negligible cost, enabling EIRES to be integrated into content-moderation pipelines, security review workflows, or on-device detection modules with minimal overhead. As a result, EIRES is scalable, hardware-efficient, and practical for large-volume or latency-sensitive applications.

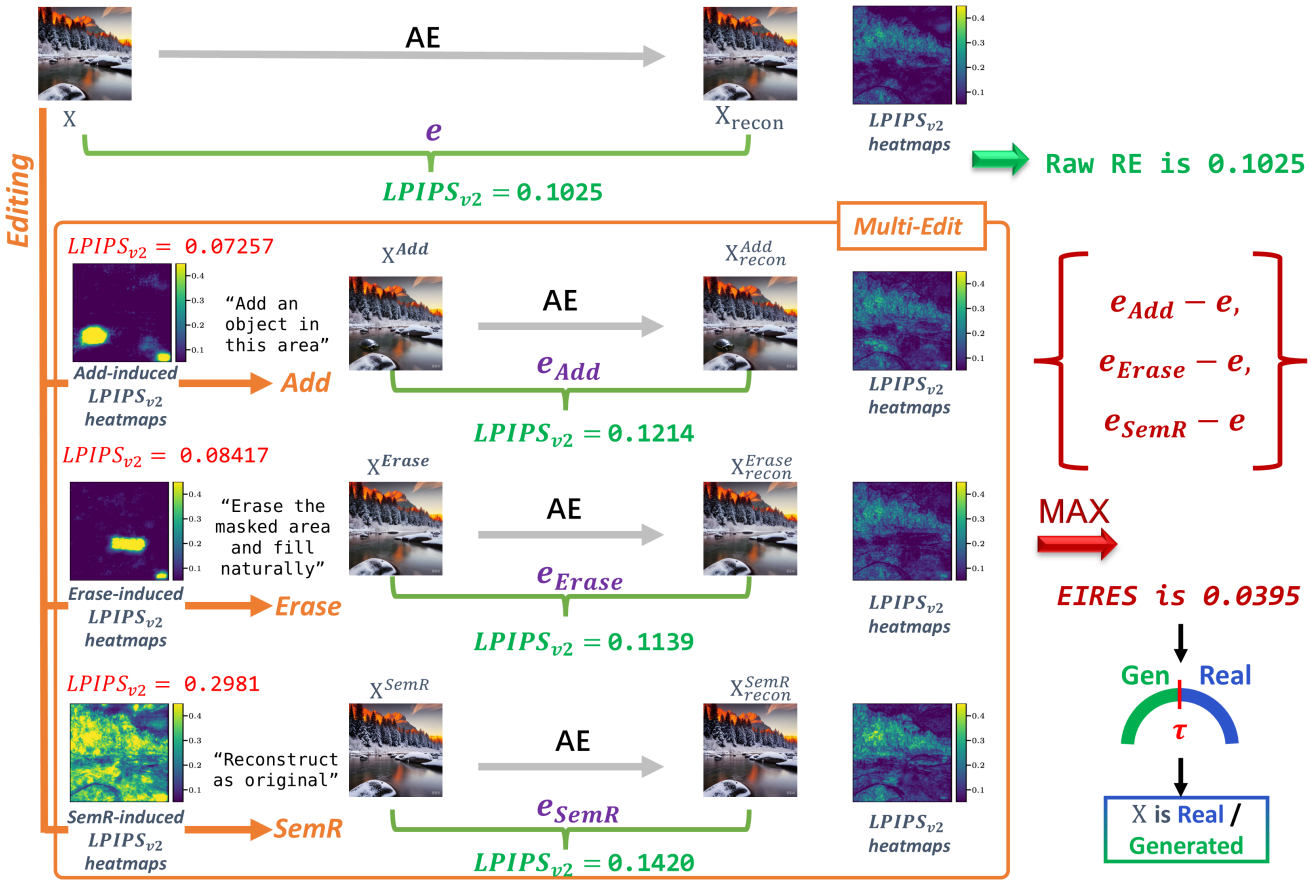


Figure 1. Example of our method for detecting an image. On the left side of the figure, we show the LPIPS distance and visual results between the original image and the edited versions induced by structured edits, highlighting the controllable nature of the edits. It is important to note that in the actual detection process, calculating these distances is not necessary. The final detection score, EIRES, is derived from the maximum deviation after applying multiple edits.

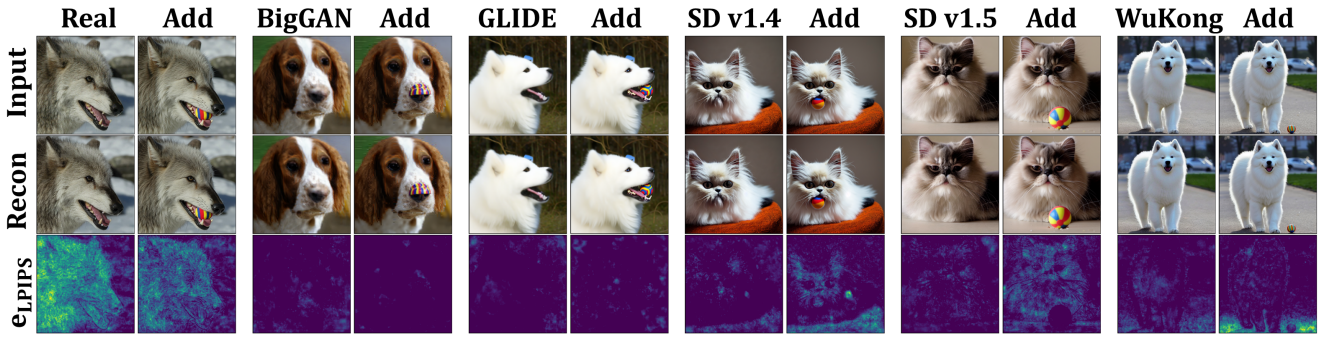
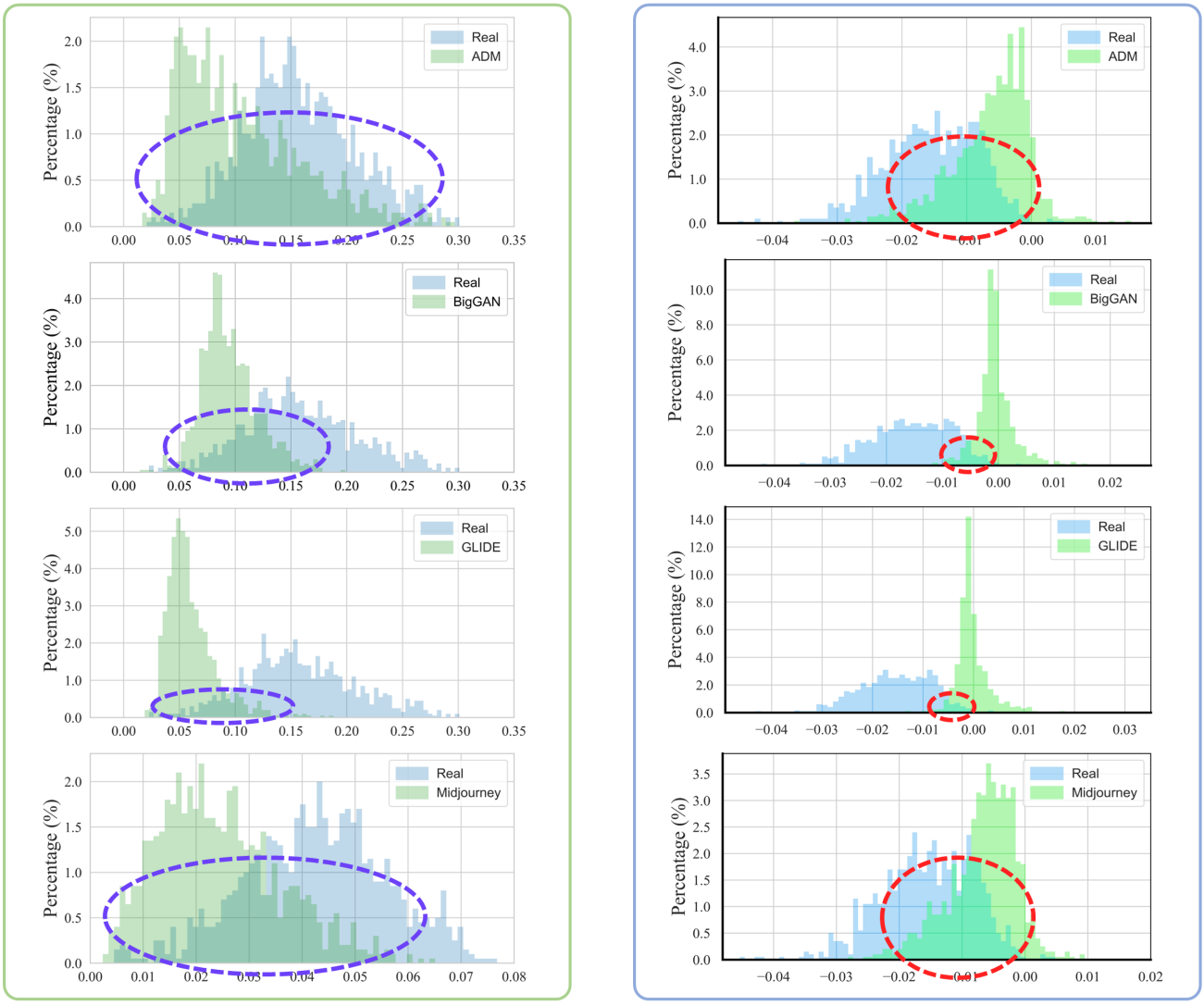


Figure 2. Visualization of reconstruction behavior under the Add editing operation across different generative models. The Add operation is implemented using Add-it [7] based on FLUX.1, with a circular mask applied to the image and the prompt "Insert a small and brightly colored ball." For each model, we show the input image, its autoencoder reconstruction, and the corresponding LPIPS heatmap of reconstruction error. Real images (leftmost) exhibit noticeably reduced reconstruction error after editing, whereas generated images (BigGAN, GLIDE, SD v1.4, SD v1.5, Wukong) display degraded or unstable reconstructions. This contrast illustrates the asymmetric edit-induced reconstruction shift that EIRES leverages for distinguishing real from generated images.



(a) Raw Reconstruction Error distribution histograms

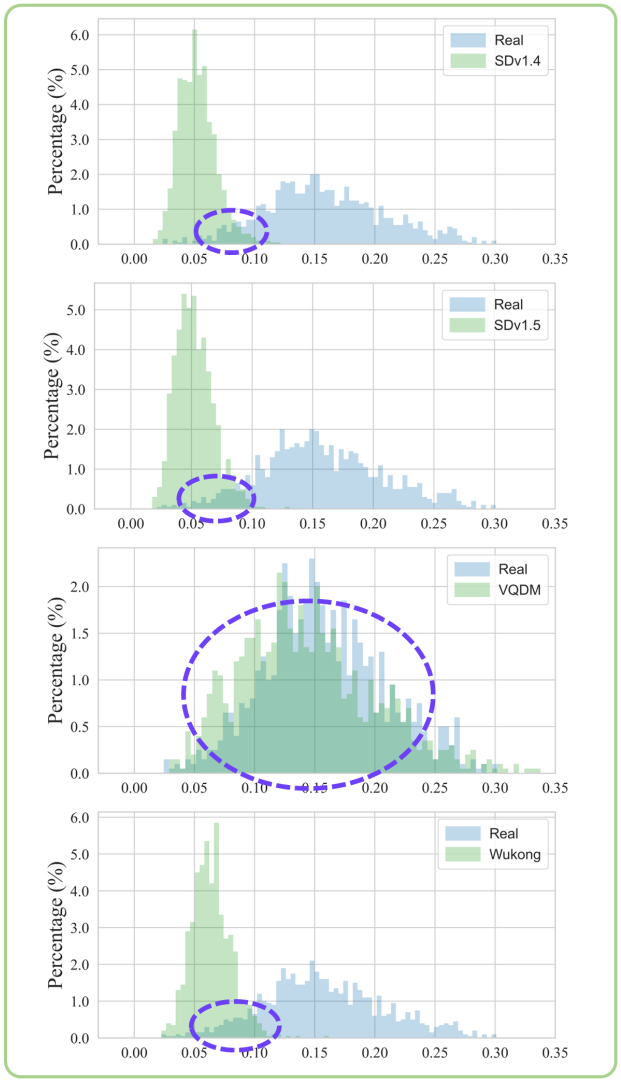
(b) Edit-Induced Reconstruction Error Shift (EIRES) distribution histograms

Figure 3. Comparison of score distributions between real and generated images across four generative models: ADM, BigGAN, GLIDE, and Midjourney. (a) Raw reconstruction error distributions show substantial overlap between real and generated images, as highlighted by the purple dashed circles, indicating limited discriminability. (b) EIRES score distributions exhibit a much clearer separation, where real and generated images form distinct clusters. The red dashed circles emphasize the regions where EIRES successfully enlarges the separation margin, revealing a significantly more discriminative signal than raw reconstruction error.

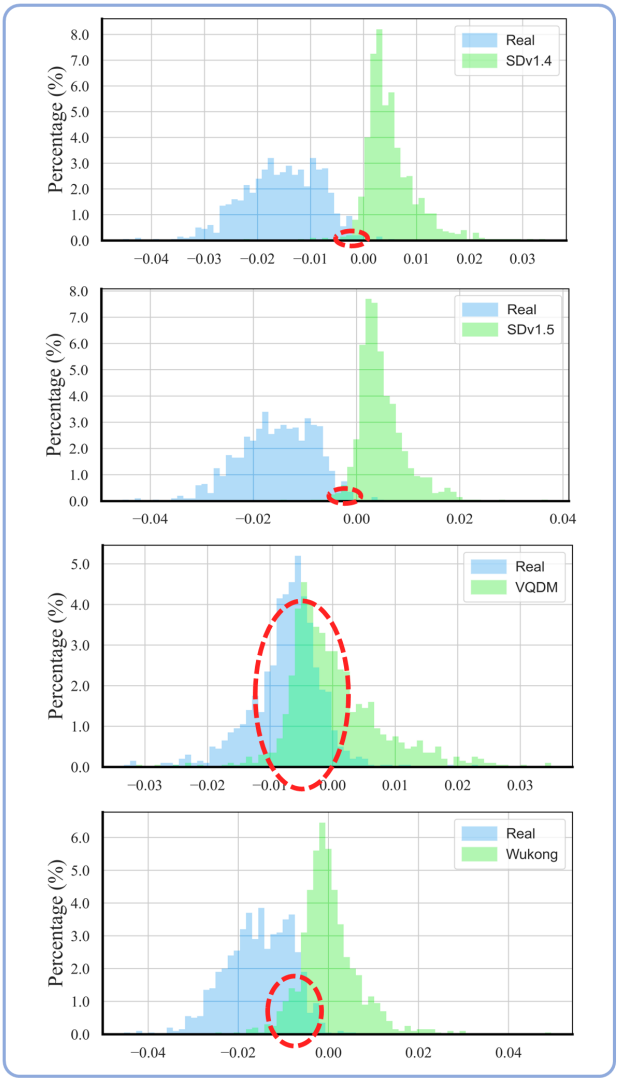
3. Additional Visual Comparisons

3.1. Edit-Induced Reconstruction Behavior

To further validate our observations, we adopt the editing strategy from the Add-it [7] framework to visualize the reconstruction behavior. As shown in Figure 2, real images show high reconstruction error before editing, especially around high-frequency textures, but this error decreases after Add editing. In contrast, generated images exhibit increased error post-editing. This indicates that real images are more responsive to structural changes, while generated ones align with the reconstruction manifold only in their original form. Interestingly, we observe that the impact of editing is not confined to the masked region. Even when the mask is very small, the reconstruction error changes across the entire image. This indicates that structured editing affects the global latent representation, altering the decoder's behavior beyond the local region. Such non-local effects support our hypothesis that dynamic reconstruction error captures broader perturbation responses, making it a more reliable signal for distinguishing real from generated images.



(a) Raw Reconstruction Error distribution histograms



(b) Edit-Induced Reconstruction Error Shift (EIRES) distribution histograms

Figure 4. Comparison of score distributions between real and generated images across four generative models: SD v1.4, SD v1.5, VQDM, and Wukong. (a) Raw reconstruction error distributions show substantial overlap between real and generated images, as highlighted by the purple dashed circles, indicating limited discriminability. (b) EIRES score distributions exhibit a much clearer separation, where real and generated images form distinct clusters. The red dashed circles emphasize the regions where EIRES successfully enlarges the separation margin, revealing a significantly more discriminative signal than raw reconstruction error.

3.2. Distribution Comparative Analysis

As shown in Figure 3 and Figure 4, we visualize the distribution of detection scores for real and generated images, where the generated images come from eight representative sources. Across all settings, EIRES consistently yields more separable distributions between real and generated samples. For instance, in the Midjourney and GLIDE settings, the real and fake distributions under raw reconstruction error exhibit significant overlap, making accurate detection difficult. In contrast, EIRES pulls the two distributions further apart, enabling clearer distinction even without any model-specific training. This advantage is especially valuable in a training-free setting, where the detector cannot rely on thresholds learned from labeled data. In many practical scenarios, labeled real or generated images may be unavailable or unreliable, especially when new generative models emerge rapidly. A method that can generalize without model-specific supervision becomes essential for

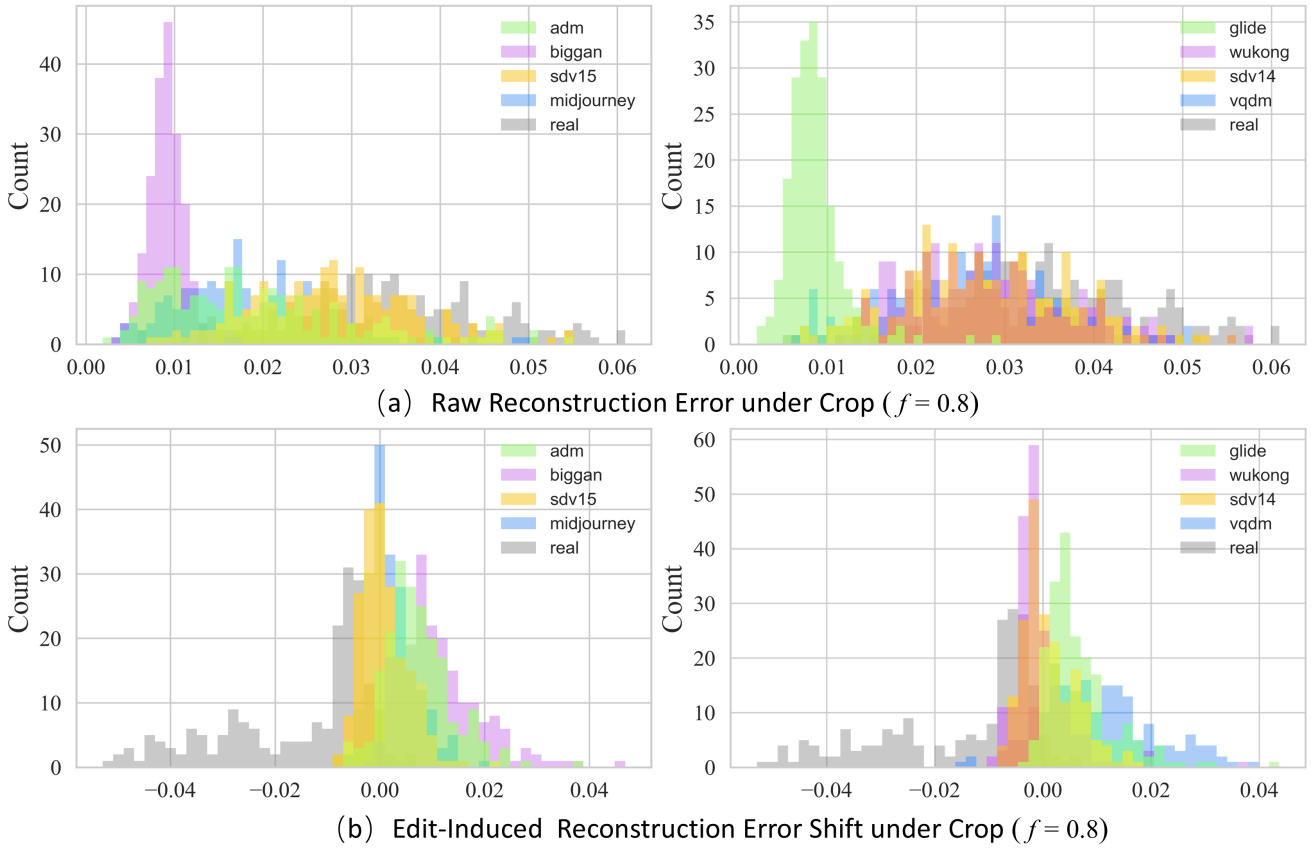


Figure 5. Distribution of detection scores under center-crop perturbation. We compare raw reconstruction error (top row) with the EIRES edit-induced reconstruction error shift (bottom row) across eight generative sources (ADM, BigGAN, VQDM, SD v1.4, SD v1.5, GLIDE, Wukong, Midjourney) under a center crop with ratio $f = 0.8$. Raw reconstruction errors exhibit substantial overlap between real and generated images, while EIRES produces more separable distributions, demonstrating improved robustness to input degradation.

scalable and robust deployment. By leveraging dynamic responses induced by structured edits, EIRES enhances the discriminative power of raw reconstruction error. Instead of passively measuring a single raw reconstruction error, EIRES actively perturbs the input through semantic edits, allowing the model to reveal how well the reconstruction aligns with the underlying manifold. This dynamic behavior not only improves separability but also mitigates overfitting to specific generators or datasets. As a result, EIRES exhibits strong generalization across diverse generative sources, making it suitable for open-world detection where the source of generated content is unknown or continuously evolving.

3.3. Distribution Comparison under Perturbations.

Figure 5 and Figure 6 compare the detection-score distributions produced by EIRES and by raw reconstruction error under two common degradations: center cropping and JPEG compression. We use center crop with ratio $f = 0.8$ and JPEG compression with $q = 80$ because these settings are standard and commonly encountered in real-world image sharing. A crop of 0.8 preserves most semantics while introducing meaningful structural changes, and JPEG quality 80 reflects typical platform compression. These moderate, realistic perturbations provide a fair and reproducible way to evaluate robustness. Across all generative models, raw reconstruction error distributions show large real–fake overlap after degradation, which collapses the margin needed for reliable thresholding. In contrast, EIRES maintains a clear separation between real and generated images even when inputs are cropped or heavily compressed. This indicates that the edit-induced reconstruction shift is remarkably stable under structural distortion and signal-level degradation. Moreover, EIRES continues to amplify intrinsic manifold differences regardless of the generator or perturbation, reinforcing the benefit of dynamic reconstruction behavior over static, single-pass error measures. These results further highlight EIRES as a practical and perturbation-robust detection mechanism suitable for deployment in real-world environments.

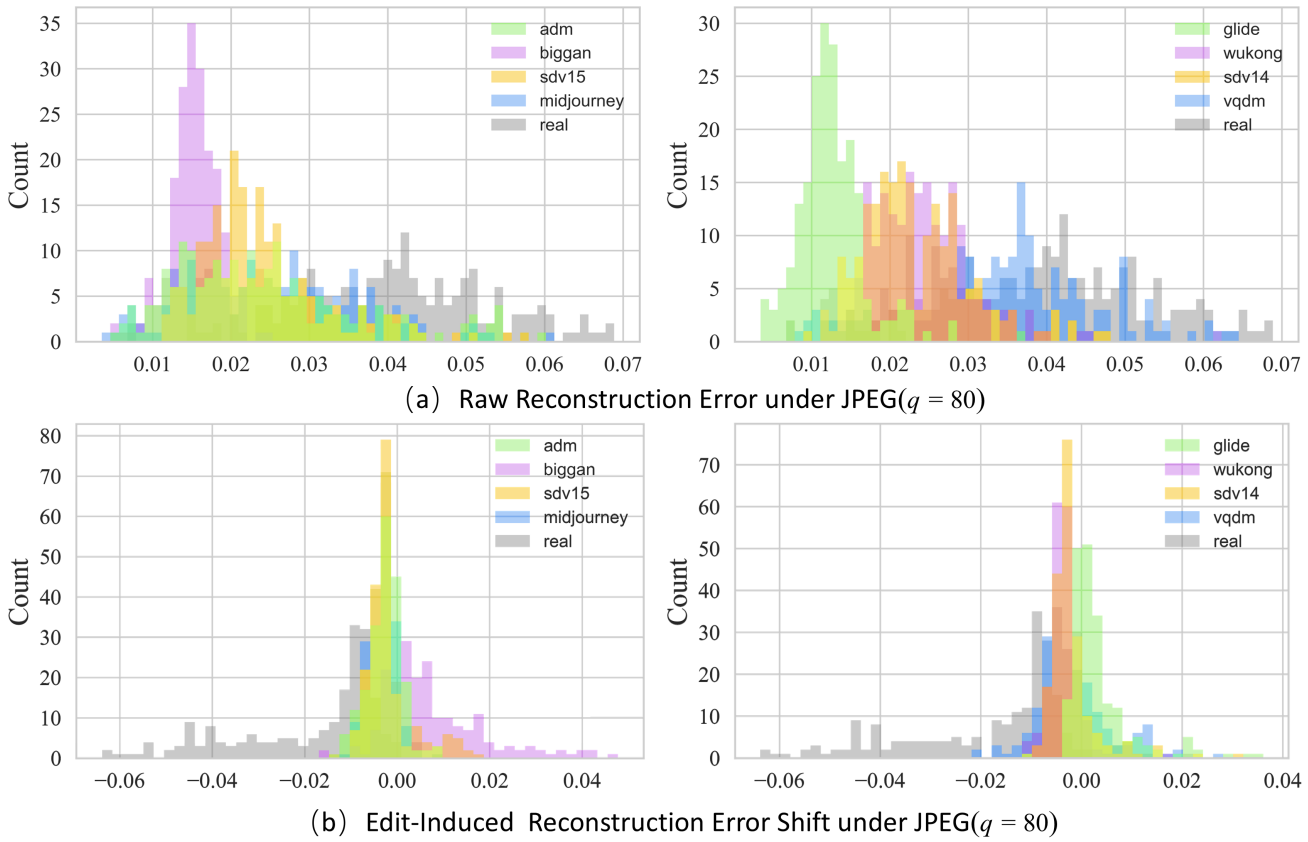


Figure 6. Distribution of detection scores under JPEG compression. We compare raw reconstruction error (top row) with the EIRES edit-induced reconstruction error shift (bottom row) for real images and generated images from eight sources under JPEG compression with quality factor $q = 80$. Across all generative models, EIRES produces significantly more separable real–fake score distributions than raw reconstruction error, demonstrating improved robustness to compression-induced degradation.

4. Proofs of Theoretical Results

To facilitate clearer understanding, we first provide a comprehensive overview of all notations used throughout the subsequent proofs, as summarized in Table 1. This notation table serves as a quick reference for the key variables, mappings, and geometric quantities involved in our analysis. Following this, we present detailed proofs for all Proposition introduced in the main paper, including the lower bound on raw reconstruction error and the lower bound under edit perturbation. These proofs offer additional mathematical clarity and support the theoretical foundations of EIRES.

4.1. Lower Bound for Reconstruction Error

We investigate the behavior of reconstruction error in the vicinity of the decoder's reconstruction manifold. Our goal is to establish a theoretical lower bound on the reconstruction error by analyzing the decoder's local geometric properties, specifically through the lens of its Jacobian condition number. In particular, we model the decoder as implicitly defining a manifold $\mathcal{M}$ in the data space. For inputs near

Symbol	Meaning
$\mathcal{M}$	Reconsrtucion manifold $\mathcal{M} = \{D(z) \mid z \in \mathbb{R}^d\}$
$\mathcal{U}$	Tubular neighbourhood of $\mathcal{M}$
x_r	Query image, off-manifold $\mathcal{M}$
$\tilde{x}_r$	Nearest point of x_r on $\mathcal{M}$, i.e., $\tilde{x}_r = \mathcal{P}(x_r)$
$T_{\tilde{x}_r}\mathcal{M}$	Tangent space of $\mathcal{M}$ at $\tilde{x}_r$
$\varepsilon_\perp$	Normal deviation $\varepsilon_\perp = x_r - \tilde{x}_r$, $\varepsilon_\perp \perp T_{\tilde{x}_r}\mathcal{M}$
z^*	Latent code of $\tilde{x}_r$: $z^* = E(\tilde{x}_r)$
$J_E(x)$	Jacobian of E at x, a $d \times HW3$ matrix
$J_D(z)$	Jacobian of D at z, a $HW3 \times d$ matrix
$\sigma_{\min}(\cdot)$	Minimum singular value
$\sigma_{\max}(\cdot)$	Maximum singular value
κ_D	condition number $\kappa_D = \sigma_{\max}(J_D)/\sigma_{\min}(J_D)$

Table 1. Symbols and notation used throughout the paper

$\mathcal{M}$, projecting onto the manifold introduces a residual primarily along the normal direction. Since the decoder exhibits limited sensitivity to off-manifold perturbations, this normal residual is expected to dominate the reconstruction error. To formalize this intuition, we first present a lemma that characterizes the local structure of the decoder around $\mathcal{M}$, thereby laying the foundation for a principled lower bound on reconstruction error.

Lemma 4.1. *Let E and D be $\mathcal{C}^1$ maps satisfying $E \circ D = \mathrm{id}$ on the data manifold $\mathcal{M}$. Then there exists an open neighbourhood $\mathcal{U}$ such that for any $x_r \in \mathcal{U}$, the nearest point on the manifold $\mathcal{M}$ is uniquely defined as*

$$\tilde{x}_r = \mathcal{P}(x_r) = \arg\min_{x \in \mathcal{M}} \|x_r - x\|_2, \tag{1}$$

and x_r can be decomposed as

$$x_r = \tilde{x}_r + \varepsilon_\perp, \;\; \varepsilon_\perp \perp T_{\tilde{x}_r}\mathcal{M}, \tag{2}$$

where $T_{\tilde{x}_r}\mathcal{M} = \mathrm{Im}\, J_D(z^*)$, $z^* := E(\tilde{x}_r)$ and $J_D(z^*)$ is the Jacobian of D evaluated at z^*.

Proof. Since D is a $\mathcal{C}^1$ map and satisfies the condition $E \circ D = \mathrm{id}$ on the latent manifold $\mathcal{M}$, the decoder D defines a regular, smooth embedding of a d-dimensional submanifold into the ambient space $\mathbb{R}^{H\times W\times 3}$. According to the classical tubular neighborhood theorem for embedded submanifolds ([6], Theorem 10.19), there exists an open neighborhood $\mathcal{U} \subset \mathbb{R}^{H\times W\times 3}$ containing $\mathcal{M}$, together with a smooth map $\mathcal{P}\colon \mathcal{U} \to \mathcal{M}$, referred to as the normal projection, such that for any point $x_r \in \mathcal{U}$, the image $\mathcal{P}(x_r) = \tilde{x}_r \in \mathcal{M}$ is the unique closest point to x_r on the manifold $\mathcal{M}$ in terms of Euclidean distance. The difference vector $\varepsilon_\perp = x_r - \mathcal{P}(x_r)$ lies in the normal bundle of $\mathcal{M}$ at the point $\tilde{x}_r$, and is therefore orthogonal to the tangent space $T_{\tilde{x}_r}\mathcal{M}$. This orthogonality condition arises from the first-order optimality condition associated with Euclidean projection onto a smooth submanifold. Moreover, the smoothness of the projection map $\mathcal{P}$ is guaranteed by the structure provided by the tubular neighborhood theorem. □

To streamline the proofs of Proposition 4.1 and Proposition 4.2, we first introduce an auxiliary lemma. This lemma provides a key geometric relation that simplifies the subsequent derivations and allows for a more structured and tractable proof.

Lemma 4.2. *Let E and D be $\mathcal{C}^1$ maps that satisfy $E \circ D = \mathrm{id}$ on the data manifold $\mathcal{M}$. Then, for any $x = D(z^*)$ with $z^* = E(x)$, it holds that*

$$\|J_E(x)\|_2 = \sigma_{\min}^{-1}\big(J_D(z^*)\big), \tag{3}$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value.

Proof. Since $E \circ D = \mathrm{id}$ on the data manifold $\mathcal{M}$, we differentiate this identity at a point $x = D(z^*)$, where $z^* = E(x)$. By the chain rule, we obtain

$$J_E(x)\, J_D(z^*) = I_d, \tag{4}$$

where I_d is the $d \times d$ identity matrix. This shows that $J_E(x)$ is a left inverse of $J_D(z^*)$.

Now recall that for any full column-rank matrix $A \in \mathbb{R}^{n\times d}$ with $n \geq d$, the Moore–Penrose pseudoinverse A^+ satisfies

$$\|A^+\|_2 = \sigma_{\min}^{-1}(A), \tag{5}$$

and furthermore achieves the minimum operator 2-norm among all left inverses of A([5], Theorem 5.2). Applying this result to $A = J_D(z^*)$, we conclude that any left inverse $J_E(x)$ satisfies

$$\|J_E(x)\|_2 \geq \|J_D(z^*)^+\|_2 = \sigma_{\min}^{-1}(J_D(z^*)). \tag{6}$$

But since $J_E(x)\, J_D(z^*) = I_d$, the bound is attained and equality holds:

$$\|J_E(x)\|_2 = \sigma_{\min}^{-1}(J_D(z^*)). \tag{7}$$

□

Based on the auxiliary lemma introduced above, we now provide the complete version of Proposition 1 from the main paper. The main text presents only a concise form of this result, and for completeness, the full statement is provided here.

Proposition 4.1 (Detailed form of proposition 1 in the main paper)**.** *Let E and D be $\mathcal{C}^1$ maps satisfying $E \circ D = \mathrm{id}$ on the data manifold $\mathcal{M}$. For any input $x \in \mathbb{R}^{H\times W\times 3}$, let $\tilde{x} \in \mathcal{M}$ denote the nearest point on the manifold, and define the normal deviation as $\varepsilon_\perp = x - \tilde{x}$. The reconstruction error behaves as follows:*

- *If $x \in \mathcal{M}$, then $\varepsilon_\perp = 0$,*

$$\|x - D(E(x))\|_2 = \mathcal{O}(\|\varepsilon_\perp\|_2^2) \to 0. \tag{8}$$

- *If $x \notin \mathcal{M}$, then*

$$\|x - D(E(x))\|_2 \geq \sqrt{1 + \kappa_D^{-2}}\,\|\varepsilon_\perp\|_2 + \mathcal{O}(\|\varepsilon_\perp\|_2^2), \tag{9}$$

where $\kappa_D = \sigma_{\max}(J_D)/\sigma_{\min}(J_D)$ is the condition number of the decoder Jacobian evaluated at $\tilde{x}$.

Proof. Let $z^* = E(\tilde{x})$ and set $\Delta z := E(x) - z^*$. A first-order Taylor expansion of E at $\tilde{x}$ gives

$$E\big(\tilde{x} + \varepsilon_\perp\big) \;=\; E(\tilde{x}) + J_E(\tilde{x})\,\varepsilon_\perp + \mathcal{O}\big(\|\varepsilon_\perp\|_2^2\big), \tag{10}$$

so that

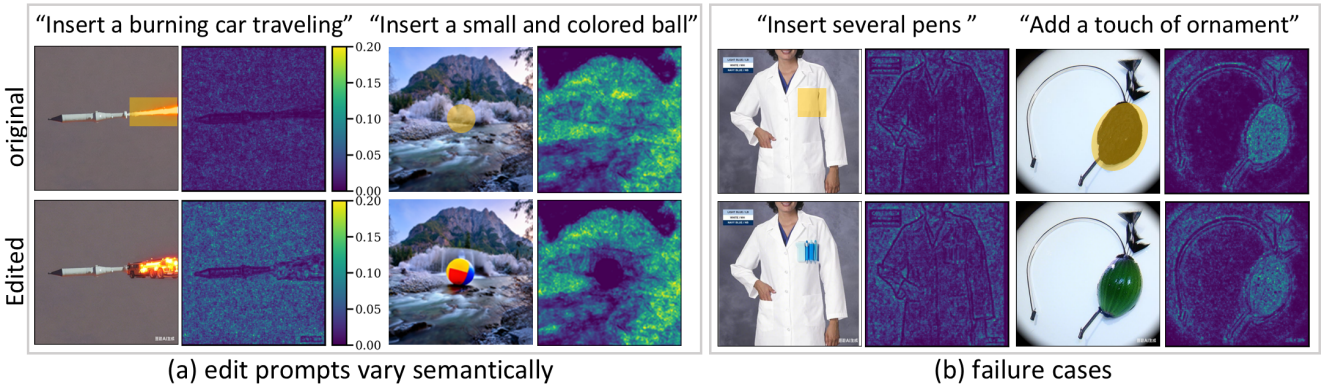


Figure 7. Further analysis of edit prompts and failure cases. (a) The method remains effective under semantically diverse edit prompts, showing opposite reconstruction responses for generated and real images. (b) Failure cases arise when real images exhibit intrinsically low reconstruction error, reducing the discriminative power of reconstruction-based signals.

$$\begin{aligned}\|\Delta z\|_2 &\geq \|J_E(\tilde{x})\,\varepsilon_\perp\|_2 - \|\mathcal{O}(\|\varepsilon_\perp\|_2^2)\|_2 \\ &\geq \sigma_{\min}(J_E(\tilde{x}))\,\|\varepsilon_\perp\|_2 + \mathcal{O}(\|\varepsilon_\perp\|_2^2). \end{aligned} \tag{11}$$

A first-order expansion of D at z^* yields

$$D(z^* + \Delta z) = D(z^*) + J_D(z^*)\,\Delta z + \mathcal{O}(\|\Delta z\|_2^2). \tag{12}$$

Hence

$$x - D(E(x)) = \varepsilon_\perp - J_D(z^*)\,\Delta z + \mathcal{O}(\|\Delta z\|_2^2). \tag{13}$$

Because $J_D(z^*)\,\Delta z \in T_{\tilde{x}}\mathcal{M}$ while $\varepsilon_\perp \perp T_{\tilde{x}}\mathcal{M}$, the first two terms in Equation (13) are orthogonal, implying:

$$\begin{aligned}\|x - D(E(x))\|_2^2 &= \|\varepsilon_\perp\|_2^2 \\ &\quad + \|J_D(z^*)\,\Delta z\|_2^2 + \mathcal{O}(\|\Delta z\|_2^3). \end{aligned} \tag{14}$$

Lemma 4.2 together with Equation (11) gives

$$\begin{aligned}\|J_D(z^*)\,\Delta z\|_2 &\geq \sigma_{\min}(J_D(z^*))\|\Delta z\|_2 \\ &\geq \kappa_D \|\varepsilon_\perp\|_2^2 + \mathcal{O}(\|\varepsilon_\perp\|_2^2). \end{aligned} \tag{15}$$

Taking square roots and absorbing high-order terms yields

$$\|x - D(E(x))\|_2 \geq \sqrt{1 + \kappa_D^{-2}}\|\varepsilon_\perp\|_2 + \mathcal{O}(\|\varepsilon_\perp\|_2^2), \tag{16}$$

then establishing Equation (9).

If $x = x_g$, i.e. $\varepsilon_\perp = 0$ and $\Delta z = 0$, immediately gives Equation (8). □

Proposition 4.1 provides strong theoretical support for using reconstruction error to distinguish real from generated images. Specifically, real images lie off the reconstruction manifold and exhibit a non-zero reconstruction error lower bounded by $\sqrt{1 + \kappa_D^{-2}}\|\varepsilon_\perp\|_2$, while generated images lie on the manifold with near-zero error. As a direct consequence of this theorem, any perturbation along the normal direction leads to a non-negligible change in reconstruction error, forming the basis for our detection method.

4.2. Lower Bound for EIRES score

Proposition 4.2 ([Proposition 2 in the main paper]). *Under the same conditions as Proposition 4.1, let $x_\mathcal{T} = x + \delta$ denote a structured edit with $\|\delta\|$ sufficiently small. Let $\tilde{x}_\mathcal{T}$ be its manifold projection and $\varepsilon_\perp^{(\mathcal{T})} = x_\mathcal{T} - \tilde{x}_\mathcal{T}$ the new normal deviation. Then:*

$$\|x_\mathcal{T} - D(E(x_\mathcal{T}))\|_2 \geq \sqrt{1 + \kappa_D^{-2}}\|\varepsilon_\perp^\mathcal{T}\|_2 + o(\|\varepsilon_\perp^\mathcal{T}\|_2^2). \tag{17}$$

Moreover, the reconstruction error shift satisfies

$$\Delta(x) = e(x_\mathcal{T}) - e(x) \propto \|\varepsilon_\perp^{(\mathcal{T})}\|_2 - \|\varepsilon_\perp\|_2. \tag{18}$$

Proof. We begin by recalling the necessary conditions from Proposition 4.1. We assume that $x_\mathcal{T} = x + \delta$ denotes a structured edit, where $\|\delta\|$ is sufficiently small. Let $\tilde{x}_\mathcal{T}$ be its manifold projection, and define $\varepsilon_\perp^{(\mathcal{T})} = x_\mathcal{T} - \tilde{x}_\mathcal{T}$ as the new normal deviation. From Lemma 4.1 and Proposition 4.1, we know that the reconstruction error satisfies:

$$\|x_\mathcal{T} - D(E(x_\mathcal{T}))\|_2 \geq \sqrt{1 + \kappa_D^{-2}}\|\varepsilon_\perp^{(\mathcal{T})}\|_2 + o(\|\varepsilon_\perp^{(\mathcal{T})}\|_2^2). \tag{19}$$

This follows from the geometric analysis and the regularity of the manifold where $D(E(x_\mathcal{T}))$ represents the reconstruction of $x_\mathcal{T}$. The term κ_D represents the curvature of the manifold at the point $x_\mathcal{T}$ and is derived from the Jacobian of the decoder, which gives the lower bound for the reconstruction error.

Now, we examine the shift in reconstruction error due to the structured edit. The shift $\Delta(x)$ in reconstruction error is given by:

$$\Delta(x) = e(x_\mathcal{T}) - e(x), \tag{20}$$

where $e(x)$ represents the reconstruction error of the image x. According to the result from Lemma 4.2 and Proposition 4.2, the shift in the reconstruction error due to the edit is proportional to the difference in the normal deviations between $x_{\mathcal{T}}$ and x:

$$\Delta(x) \propto \|\varepsilon_{\perp}^{(\mathcal{T})}\|_2 - \|\varepsilon_{\perp}\|_2. \tag{21}$$

This result directly follows from the fact that the reconstruction error shift depends on how much the image deviates from the manifold after the edit. The term $\varepsilon_{\perp}$ represents the normal deviation of the original image x, while $\varepsilon_{\perp}^{(\mathcal{T})}$ represents the normal deviation of the edited image $x_{\mathcal{T}}$.

Combining these results, we conclude that:

$$\|x_{\mathcal{T}} - D(E(x_{\mathcal{T}}))\|_2 \geq \sqrt{1+\kappa_D^{-2}}\|\varepsilon_{\perp}^{(\mathcal{T})}\|_2 + o(\|\varepsilon_{\perp}^{(\mathcal{T})}\|_2^2), \tag{22}$$

and the reconstruction error shift satisfies:

$$\Delta(x) \propto \|\varepsilon_{\perp}^{(\mathcal{T})}\|_2 - \|\varepsilon_{\perp}\|_2. \tag{23}$$

□

Proposition 4.2 provides a critical insight into how the reconstruction error behaves under structured edits. Specifically, it shows that the deviation of the edited image from the manifold, denoted as $\varepsilon_{\perp}^{(\mathcal{T})}$, plays a central role in determining the magnitude of the reconstruction error shift. This proposition emphasizes that the reconstruction error of the edited image $x_{\mathcal{T}}$ can be bounded in terms of the normal deviation from the manifold, which helps us understand the behavior of the reconstruction error in a geometrically rigorous manner. The relationship between $\|\varepsilon_{\perp}^{(\mathcal{T})}\|_2$ and the reconstruction error shift is crucial for establishing the lower bound and understanding the sensitivity of the model to structured edits.

5. Limitations and Future Work

While EIRES demonstrates strong separability and robustness across diverse generative models, several limitations remain. The method relies on the assumption that real and generated images exhibit different responses to structured editing. This assumption holds in most cases, even when the edit prompts are not strictly semantically consistent, as illustrated in Figure 7 (a). However, the response gap may diminish in more challenging scenarios, such as heavily stylized content, severe domain shifts, or real images that are inherently easy to reconstruct. In particular, failure cases arise when real images exhibit unusually low reconstruction error under the reconstruction model and therefore resemble generated images. This effect weakens detection signals that rely on reconstruction changes, as shown in Figure 7 (b). These observations suggest that the effectiveness of edit-induced probing is influenced by the intrinsic reconstruction characteristics of the input.

Improving robustness under such conditions may require adaptive editing strategies that dynamically adjust perturbation strength, edit type, or spatial region based on the input image. Moreover, EIRES currently focuses on image-level detection. Extending the edit-induced reconstruction paradigm to tasks such as localizing manipulated regions, analyzing multimodal consistency, or detecting generated videos represents a promising direction. Exploring these directions may further enhance the generality and forensic value of edit-induced probing signals.

References

[1] Doubao large model platform. https://www.doubao.com. Accessed: 2025-01-20. 1

[2] Midjourney ai image generation and editing. https://www.midjourney.com. Accessed: 2025-01-20. 1

[3] Baidu Inc. Ernie-irag-edit: Baidu's image editing model. https://ai.baidu.com/model/ernie-irag-edit, 2025. Accessed: 2025-08-02. 1

[4] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 1

[5] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013. 7

[6] John M Lee. *Introduction to Smooth Manifolds*. Springer, 2003. 7

[7] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3