A. Theoretical Analysis

A.1. Is it always possible to distinguish between Generated content and real images?

The recent work explores the detection of AI-generated content by analyzing the AUROC for any detector D. It leverages Le Cam's lemma [22, 45], which states that for any distributions G and H, given an observation s, the minimum sum of Type-I and Type-II error probabilities in testing whether $s \sim G$ or $s \sim H$ is equal to $1 - d_{\text{TV}}(G, H)$, where d_{TV} denotes the total variation distance between the two distributions. This result can be interpreted as:

$$\text{TPR}_\gamma \leq \min\{\text{FPR}_\gamma + d_{\text{TV}}(G, H), 1\}, \tag{5}$$

where $\text{TPR}_\gamma \in [0, 1]$. The upper bound in equation 5 is leveraged in one of the recent work [36] to derive AUROC upper bound $\texttt{AUC} \leq \frac{1}{2} + d_{\text{TV}}(G, H) - \frac{d_{\text{TV}}(G,H)^2}{2}$ which holds for any D. This upper bound led to the claim of impossibility results for reliable detection of AI-Generated content when $d_{\text{TV}}(G, H)$ is approaching 0. The upper bound in equation 5 is also interpreted as either certain real images will be detected falsely as AI-generated content will not be detected reliably when $d_{\text{TV}}(G, H)$ is small. However, as discussed in Sec. 3, the Likelihood-Gap Hypothesis guarantees that the difference between the two distributions is significant enough ($d_{\text{TV}}(G, H)$ or $d_{\text{KL}}(G, H)$ is greater than some positive gap). This implies it is always possible to distinguish between real and machines.

A.2. Principled Choice of K

In Sec. 3, we propose the **Likelihood-Gap Hypothesis**, which posits that the expected log-likelihood of the machine generation process G exceeds that of the human generation process H by a positive gap, $\Delta > 0$. To exploit this difference between the distributions, we introduce a distance function $D(Y, Y')$ that quantifies the similarity between two images Y and Y'. This distance function can also be interpreted as a kernel function used in kernel density estimation.

By re-prompting the masked pixels, we can evaluate how closely the remaining pixels Y_0 align with the machine-generated distribution: $\hat{D}(Y_0, \{Y_k\}_{k\in[K]}) := \frac{1}{K}\sum_{k=1}^{K} D(Y_0, Y_k)$, where K is the number of times of re-prompting.

Similar to the kernel density estimation, we can use this quantity and some threshold to determine whether to accept or reject that $S \sim G$. Under certain assumptions, this estimator enjoys $n^{-1/2}$-consistency via Hoeffding's argument. In the following, we provide a formal argument.

Assumption 1 *Suppose we have a given human-generated content $[X, Y_0] \in \text{supp}(h)$ and a machine-generated remaining pixels $\tilde{Y}_0$, consider the random variable $D(Y_0, Y')$ and where Y' is sampled by re-prompting given X, that is $Y' \sim G(\cdot|X)$. We assume $D(Y_0, Y')$ and $D(\tilde{Y}_0, Y')$ are σ-sub-Gaussian. We also assume that the distance gap is significant: $\mathcal{E}_{Y'\sim G}[D(Y_0, Y')|X] - \mathcal{E}_{Y'\sim G}[D(\tilde{Y}_0, Y')|X] > \Delta$.*

From this assumption, we can derive that it suffices to re-prompt $\Omega\big(\frac{\sigma \log(1/\delta)}{\Delta^2}\big)$ times.

Proof Note that $\mathcal{E}[\hat{D}] = \mathcal{E}[D]$ and the distribution is sub-Gaussian. By Hoeffding's inequality, we have that with probability at least $1 - \delta$,

$$\left|\frac{1}{K}\sum_{k=1}^{K} D(Y_0, Y_k) - \mathcal{E}_{Y'\sim G}[D(Y_0, Y')|X]\right| \leq \sqrt{\frac{\sigma \log(\delta/2)}{K}}.$$

Similarly, we have that with probability at least $1 - \delta$,

$$\left|\frac{1}{K}\sum_{k=1}^{K} D(\tilde{Y}_0, Y_k) - \mathcal{E}_{Y'\sim G}[D(\tilde{Y}_0, Y')|X]\right| \leq \sqrt{\frac{\sigma \log(\delta/2)}{K}}.$$

By the union bound, we have that with probability $1 - 2\delta$,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} D(Y_0, Y_k) - \frac{1}{K}\sum_{k=1}^{K} D(Y_0, Y_k) \\
&> \frac{1}{K}\sum_{k=1}^{K} D(Y_0, Y_k) - \mathcal{E}_{Y'\sim G}[D(\tilde{Y}_0, Y')|X] \\
&\quad - \frac{1}{K}\sum_{k=1}^{K} D(\tilde{Y}_0, Y_k) + \mathcal{E}_{Y'\sim G}[D(\tilde{Y}_0, Y')|X] + \Delta \\
&\geq \Delta - 2\sqrt{\frac{\sigma \log(\delta/2)}{K}}.
\end{aligned}$$

If we set $K = \Omega\big(\frac{\sigma \log(1/\delta)}{\Delta^2}\big)$, then there is a gap between the real distance and the machine's distance.

B. More Details of Scoring Function δ

In this section, we provide additional details about scoring function δ including PSNR, SSIM, L1 distance and L2 distance. Let $\mathbf{I} \in \mathbb{R}^{h\times w\times c}$ be the original image, where h and w are the height and width, respectively, and c is the number of channels. Let $\mathbf{I}' \in \mathbb{R}^{h\times w\times c}$ be the recovered image. Let MAX denote the maximum possible pixel value (e.g., 255 for 8-bit images).

Peak Signal-to-Noise Ratio (PSNR) measures the ratio between the maximum possible value of a pixel and the power of the distortion (i.e., Mean Squared Error) between the original and reconstructed images. The Mean Squared Error (MSE) defined as follows:

$$\text{MSE}(\mathbf{I}, \mathbf{I}') = \frac{1}{h \times w \times c}\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c}\left(\mathbf{I}(i, j, k) - \mathbf{I}'(i, j, k)\right)^2.$$

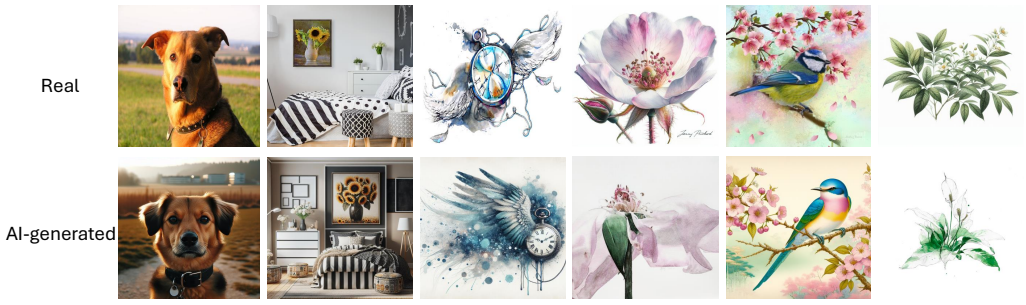

Figure 6. Examples of hard cases for distinction in human evaluations.

The PSNR formula: $\text{PSNR}(\mathbf{I}, \mathbf{I}') = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}(\mathbf{I},\mathbf{I}')}\right)$, a higher PSNR value indicates a smaller difference between the images, implying better recovery.

Structural Similarity Index (SSIM) is designed to measure perceptual differences between two images, taking into account luminance, contrast, and structural information. The formula is: $\text{SSIM}(\mathbf{I}, \mathbf{I}') = \frac{(2\mu_{\mathbf{I}}\mu_{\mathbf{I}'}+C_1)(2\sigma_{\mathbf{II}'}+C_2)}{(\mu_{\mathbf{I}}^2+\mu_{\mathbf{I}'}^2+C_1)(\sigma_{\mathbf{I}}^2+\sigma_{\mathbf{I}'}^2+C_2)}$, where$\mu_{\mathbf{I}}$ and $\mu_{\mathbf{I}'}$ are the means of $\mathbf{I}$ and $\mathbf{I}'$. $\sigma_{\mathbf{I}}^2$ and $\sigma_{\mathbf{I}'}^2$ are the variances of $\mathbf{I}$ and $\mathbf{I}'$. $\sigma_{\mathbf{II}'}$ is the covariance between $\mathbf{I}$ and $\mathbf{I}'$. C_1 and C_2 are small constants to stabilize the division. The SSIM values range from -1 to 1, where 1 indicates a perfect match.

L1 distance measures the absolute difference between corresponding pixels of original and reconstructed images: $L_1(\mathbf{I}, \mathbf{I}') = \frac{1}{h\times w\times c}\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c}|\mathbf{I}(i,j,k) - \mathbf{I}'(i,j,k)|$, where a lower L1 value indicates a smaller difference between the images.

L2 distance measures the squared difference between the corresponding pixels of the original and reconstructed images: $L_2(\mathbf{I}, \mathbf{I}') = \frac{1}{h\times w\times c}\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c}(\mathbf{I}(i,j,k) - \mathbf{I}'(i,j,k))^2$, where a lower L2 value indicates a smaller difference between the images. L2 distance is related to PSNR as it forms the basis of its calculation.

C. Additional Experimental details

We provide a detailed description of the datasets and model used in this work:

Stable Diffusion [33] is a text-to-image model based on diffusion techniques. Originating from latent diffusion, its model and weights have been publicly released. Stable Diffusion was trained on pairs of images and captions from LAION-5B[38], an open large-scale dataset for training image-text models.

Guided Diffusion [12] is a diffusion model that uses gradients from a classifier to guide the denoising process during image synthesis. This approach has proven effective for image generation, surpassing GANs in terms of fidelity while maintaining broad distribution coverage.

GLIDE [27] is a text-guided diffusion model designed for photorealistic image generation and editing. It employs classifier-free guidance to enhance image quality while maintaining fidelity to text prompts.

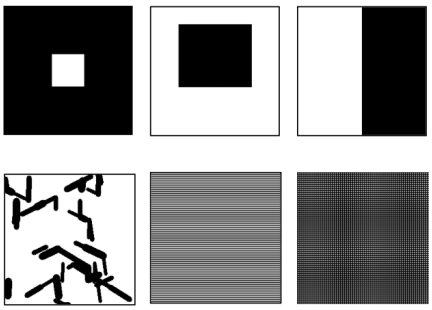

Figure 7. Example of different masks.

LDM [33] apply diffusion processes in the latent space of pretrained autoencoders rather than directly in high-dimensional pixel space. This approach significantly reduces computational costs while retaining high-quality image synthesis.

DALL-E [30] is an advanced generative model developed by OpenAI for text-to-image synthesis. It creates highly detailed and imaginative images from natural language descriptions, demonstrating strong performance in generating diverse and realistic visuals while enabling creative applications in content generation and design.

DALL-E 3 [3] is the latest version of OpenAI's text-to-image generative model, offering significant improvements in fidelity, creativity, and alignment with text prompts. It sets a new standard in text-to-image synthesis.

Hardware and software. Our framework was implemented using PyTorch 2.3.1. Experiments are performed using the RTX A6000.

D. Visualization of Different Masks

This work supports flexibility with various types of masks. Figure 7 illustrates some examples of different masks.

E. Hard Examples for Distinguishing

Table 6 demonstrates examples where human struggle to accurately differentiate real from AI-generated images.