1 Background

The development of diffusion models has ushered in an era of eerily realistic generated content across a range of media, such as images, audio, and text. Although the capabilities of these models are impressive, it is important to reflect on their future implications. As these models continue to improve, they may pose significant societal risks, such as enabling the spread of misinformation and making copyright enforcement more difficult. As a result, it is crucial to have reliable methods for identifying artificially generated content. In this project, we will explore the integration of watermarking techniques into the reverse diffusion process, so that the presence of a watermark serves as a strong statistical indicator that the content originated from an AI system.

Due to the prevalence and strong capabilities of recent image diffusion models, we will focus our project on watermarking images. Diffusion models work by learning to reverse an iterative noise-adding procedure known as the "forward process." Beginning with a clean image, the forward process iteratively adds Gaussian noise according to a parameterized noise schedule, eventually turning the image into pure Gaussian noise [3]. During inference, the model performs the "reverse process", iteratively predicting and removing the noise added in previous steps. This reverse process can take place either directly in image space or, as is increasingly common, in a latent space [3, 4]. We hope to investigate both areas, studying how watermarking in specific color channels (e.g. RGB and YUV) and transform domains (e.g. DFT, DCT, and DWT) compares to watermarking directly in the latent space. Thus, our **research question** is the following:

How and in what domain (e.g. image-space, latent-space, discrete transforms) should a watermark be embedded in the reverse process of an image diffusion model to achieve the best balance between imperceptibility and robustness?

2 Application Survey

The traditional approach to image watermarking involves embedding signals in the spatial domain by directly altering pixel values, or in a frequency domain by modifying transform coefficients such as those from the Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), or Discrete Wavelet Transform (DWT) [2]. These methods have been widely applied for copyright protection and tamper detection, but they are becoming increasingly inadequate for identifying AI-generated images, where conventional overlays or post-processing watermarks can be easily removed or forged.

Recent work has shifted toward watermarking within generative AI models. Instead of directly modifying pixels, these approaches embed watermarks in the latent space of the model by altering the predicted noise during the reverse diffusion process [1]. Current evaluations of watermarking methods focus on two criteria: imperceptibility, measured by image quality metrics such as Peak

Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), and robustness, which refers to the watermark's ability to survive distortions such as compression, cropping, and noise injection.

In this project, we will hold the overall setup constant while varying only the placement of the watermark. This will allow us to measure how embedding location affects imperceptibility and robustness, and to understand why certain domains offer more favorable trade-offs.

3 Information Theory Perspective

From an information-theoretic perspective, watermarking can be modeled as a communication problem over a noisy channel. We want to guarantee reliable communication of the embedded watermark across a family of adversarial channels, subject to imperceptibility constraints. More specifically, we seek to minimize the probability of decoding error P_ϵ over a family of channels $\{C_i\} \subseteq \mathcal{A}$, where $\mathcal{A}$ denotes the set of adversarial channels:

$$\sup_{C_i \in \mathcal{A}} P_\epsilon(C_i) < \delta$$

for some small δ.

In addition to robustness, watermarking must also satisfy imperceptibility, which can be modeled as a distortion constraint. Unlike classical channel coding, where the goal is just to maximize rate, watermarking creates additional constraints: the distortion it introduces must fit within a perceptual budget (hidden from humans):

$$\mathbb{E}[d(X, Y)] \leq D_{max}$$

where $d(\cdot)$ is the distortion measure, Y is the watermarked signal (image), and X is the original. Since human perception is context-dependent, distortion measures are often adaptive. In practice, imperceptibility is usually evaluated via image quality metrics such as PSNR and SSIM. Overall, this makes watermarking a joint rate–distortion and channel coding problem.

4 Conclusion

Our project aims to investigate how a watermark can be embedded in the reverse process of image diffusion models. We will examine different domains in which the watermark can be inserted, considering both image-space and latent-space approaches. To evaluate each method, we will measure robustness using standard image perturbation techniques (e.g., geometric transformations) and assess imperceptibility using metrics such as PSNR. Finally, we will explore information-theoretic justifications for why certain color channels or transforms may perform better than others.

References

[1] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Universal watermarking for arbitrary resolution images. *arXiv preprint arXiv:2311.18297*, 2023.

[2] Lele Cao. Watermarking for ai content detection: A review on text, visual, and audio modalities. In *Proceedings of the 1st Workshop on GenAI Watermarking, collocated with ICLR*, Stockholm, Sweden, 2025. arXiv preprint arXiv:2504.03765.

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[4] Li Wang, Boyan Gao, Yanran Li, Zhao Wang, Xiaosong Yang, David A. Clifton, and Jun Xiao. Exploring the latent space of diffusion models directly through singular value decomposition, 2025.