SUPPLEMENTARY FILE
FLASHEDIT: DECOUPLING SPEED, STRUCTURE, AND SEMANTICS FOR PRECISE IMAGE EDITING

Anonymous authors
Paper under double-blind review

1 A COMPREHENSIVE OVERVIEW OF DIFFUSION MODELS

Positioned at the forefront of generative modeling, Diffusion Probabilistic Models (DMs) (Ho et al. (2020); Nichol & Dhariwal (2021); Kingma et al. (2021); Karras et al. (2022); Song et al. (2021); Chen et al. (2023); Salimans & Ho (2022)) represent a significant paradigm shift, celebrated for their exceptional capacity to generate high-fidelity outputs. Unlike Generative Adversarial Networks (GANs), which often suffer from training instability, or Variational Autoencoders (VAEs), which can produce blurry results, DMs offer stable training and unparalleled sample quality. Their influence spans numerous applications, from image and video creation to text-to-image synthesis and audio generation. By providing a powerful, principled methodology for learning complex data distributions, these models have become a fundamental pillar of contemporary artificial intelligence. More recently, the fusion of diffusion principles with Transformer architectures has given rise to Diffusion Transformers (DiTs) (Yang et al. (2025); Peebles & Xie (2023); Labs (2024); Hatamizadeh et al. (2024); Lin et al. (2022); Kim et al. (2024)), which demonstrate remarkable scaling properties. This section offers a thorough examination of their core concepts, mathematical underpinnings, and architectural variations.

1.1 CORE PRINCIPLES OF DIFFUSION MODELS

Conceptually rooted in non-equilibrium thermodynamics and statistical mechanics, DMs are a class of latent variable models. They operate through a carefully defined two-stage process that mirrors the physical phenomenon of diffusion. The first stage, the **forward process**, systematically degrades structured data (a low-entropy state) by incrementally adding Gaussian noise over a series of timesteps until only pure, unstructured noise remains (a high-entropy state). The second stage, the **reverse process**, learns to invert this procedure. Here, a neural network is trained to methodically remove the noise, starting from a random sample and progressively refining it to reconstruct a sample from the original data distribution, effectively decreasing the system's entropy. This approach transforms the intractable problem of modeling a complex data manifold into a more manageable sequence of denoising steps. The power of DMs originates from this iterative refinement, which enables the generation of highly detailed and coherent samples.

1.2 THE FORWARD NOISING PROCESS

The forward process Ho et al. (2020) is a fixed (non-learned) Markov chain that gradually perturbs an initial data sample $x_0 \sim q(x_0)$ over T discrete timesteps. At each step t, a small amount of Gaussian noise is added according to a variance schedule β_t:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \tag{1}$$

The sequence $\{\beta_t\}_{t=1}^T$ is a critical hyperparameter, typically chosen such that $\beta_1 < \beta_2 < \cdots < \beta_T$. Common schedules include linear, quadratic, or cosine schedules, which control the rate of information destruction. As $t \to T$, the cumulative effect of the noise addition ensures that x_T converges to an isotropic Gaussian distribution, $x_T \sim \mathcal{N}(0, I)$.

A key property of this process is that we can sample x_t directly from x_0 at any timestep t without iteration. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. Using the reparameterization trick, we can write x_t as a function of x_0 and a random noise variable $\epsilon \sim \mathcal{N}(0, I)$:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{2}$$

This gives the closed-form sampling distribution:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I). \tag{3}$$

Here, $\sqrt{\bar{\alpha}_t}$ represents the signal rate, while $\sqrt{1-\bar{\alpha}_t}$ is the noise rate. This property is instrumental for efficient model training.

1.3 The Learned Reverse Denoising Process

The reverse process is the generative core of the model, tasked with inverting the forward noising process. It begins with a sample from the prior distribution, $x_T \sim \mathcal{N}(0, I)$, and aims to learn the true posterior distribution $q(x_{t-1}|x_t)$ to iteratively denoise the sample back to x_0. This reverse process is modeled as a Markov chain with learned parameters θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{4}$$

While the true posterior $q(x_{t-1}|x_t)$ is intractable, its form conditioned on the original data x_0 is tractable and Gaussian:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \tag{5}$$

with mean and variance given by:

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \tag{6}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{7}$$

The goal of the neural network $\mu_\theta(x_t, t)$ is to approximate $\tilde{\mu}_t(x_t, x_0)$. By substituting x_0 from Eq. 2 into the expression for $\tilde{\mu}_t$, it can be shown that predicting the mean $\tilde{\mu}_t$ is mathematically equivalent to predicting the noise ϵ that was added to x_0 to produce x_t. This insight leads to a simplified and highly effective training objective Ho et al. (2020). Instead of optimizing the full variational lower bound (VLB), we can directly minimize the mean squared error between the true and predicted noise:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right], \tag{8}$$

where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, and ϵ_θ is the output of the neural network. This objective is not only computationally efficient but has also been shown to produce superior results.

1.4 Architectural Foundations

1.4.1 U-Net Backbone

The prevalent architectural choice for many diffusion models, especially in the image domain, is the U-Net Rombach et al. (2022); Prasad et al. (2023); Xing et al. (2023). Its encoder-decoder structure with skip connections is highly effective for tasks requiring pixel-level prediction. The network takes two primary inputs: the noisy data x_t and the current timestep t. The timestep t is crucial context, as it informs the network about the level of noise it needs to remove. It is typically encoded using sinusoidal embeddings, inspired by Transformer positional encodings, and then incorporated into the network's intermediate layers. A key enhancement in modern U-Nets is the integration of **self-attention** blocks, usually at lower-resolution feature maps, which enable the model to capture long-range spatial dependencies. For **conditional generation** (e.g., text-to-image), conditioning information c (like text embeddings from a CLIP model) is integrated into the U-Net, often via cross-attention mechanisms, allowing external signals to guide the denoising process.

1.4.2 Diffusion Transformers (DiT)

A more recent innovation is the Diffusion Transformer (DiT) (Peebles & Xie (2023)), which replaces the convolutional U-Net with a Transformer backbone. This approach treats the diffusion process as a sequence modeling problem. A noisy image x_t is first broken down into a sequence of non-overlapping patches, which are then linearly projected to form input tokens. These tokens, along with timestep and conditional embeddings, are processed by a series of Transformer blocks. The final output tokens are decoded to predict the output noise for each patch. The primary advantage of the DiT architecture is its demonstrated **scalability**; its performance improves predictably and significantly with increased model size and computational resources, setting new benchmarks in high-fidelity image generation.

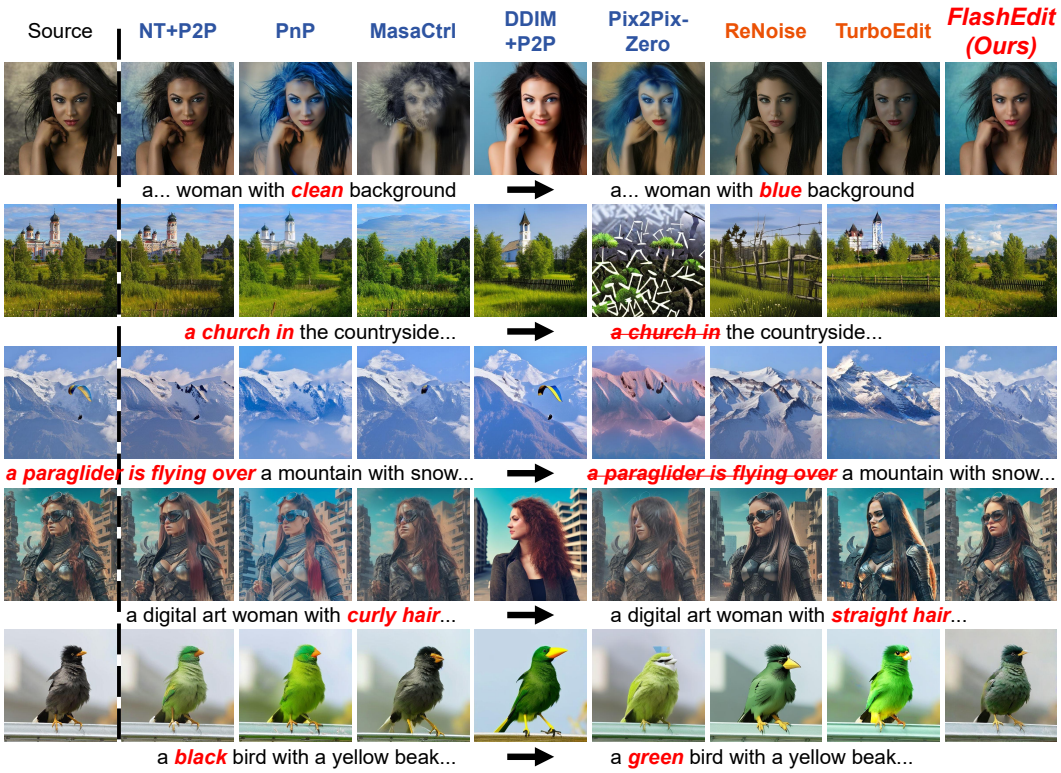

Figure 1: More comparative visual results.

2 Evaluation Metrics

To quantitatively evaluate our method, we utilize a comprehensive suite of metrics designed to assess different facets of image quality, including pixel-level accuracy, perceptual similarity, and semantic consistency with the text prompt.

2.1 Pixel-Level Fidelity Metrics

These metrics measure the exactness of the reconstruction at the pixel level by directly comparing the pixel values between the original and edited images.

- **Mean Squared Error (MSE)** computes the average squared difference between the pixel values of two images. It provides a strict, quantitative measure of reconstruction error. A **lower** MSE value signifies a smaller pixel-wise deviation and thus a higher-fidelity reconstruction.
- **Peak Signal-to-Noise Ratio (PSNR)** (Huynh-Thu & Ghanbari (2008)) is derived from the MSE and represents the ratio between the maximum possible pixel value (signal) and the magnitude of the reconstruction error (noise). It is expressed on a logarithmic scale (dB). A **higher** PSNR value indicates a higher quality reconstruction with less error.

2.2 Perceptual Similarity Metrics

These metrics aim to approximate human perception of image similarity more closely than pixel-level metrics.

- **Structural Similarity Index Measure (SSIM)** (Wang et al. (2004)) evaluates image similarity by comparing three key features: luminance, contrast, and structure. It is designed to be more robust to minor pixel variations and better align with how humans perceive visual

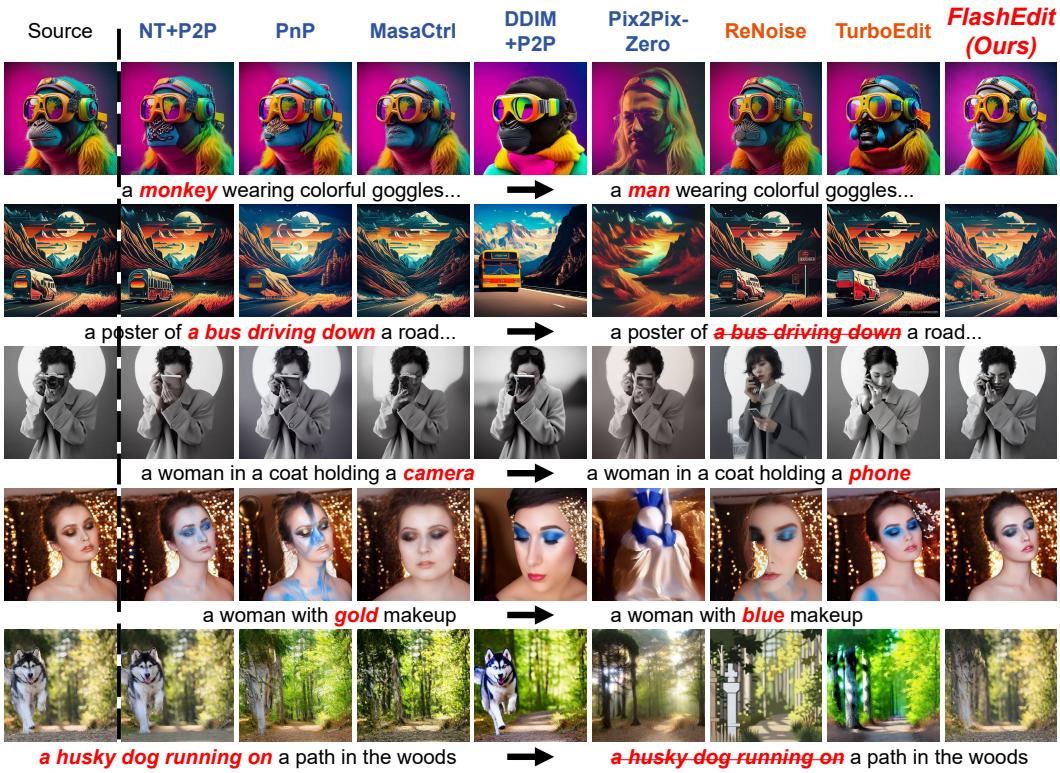

Figure 2: More comparative visual results.

quality. SSIM values range from -1 to 1, where a **higher** value closer to 1 indicates greater perceptual similarity.

- **Learned Perceptual Image Patch Similarity (LPIPS)** (Zhang et al. (2018)) measures the perceptual distance between two images using features extracted from a pre-trained deep neural network (e.g., AlexNet or VGG). By comparing images in a feature space that is more aligned with human vision, LPIPS has shown a strong correlation with human judgment of image similarity. A **lower** LPIPS score indicates that two images are more perceptually similar.

2.3 Semantic Alignment Metric

This metric assesses how well the generated or edited image aligns with the meaning of the input text prompt.

- **CLIP Score** (Radford et al. (2021)) measures the semantic correspondence between an image and a text description using a pre-trained CLIP (Contrastive Language-Image Pre-training) model. It calculates the cosine similarity between the image and text embeddings in a shared multimodal space. A **higher** CLIP Score indicates a stronger semantic alignment between the image content and the textual instruction.

3 More Visual Experiments

To provide further evidence of our method's efficacy and robustness, we present a comprehensive collection of additional qualitative results in Figure 1, Figure 2 and Figure 3. This section showcases the performance of **FlashEdit** across a diverse array of text-guided image editing tasks, including attribute modification, object deletion, and complex scene manipulation. We compare our results against a range of state-of-the-art multi-step and few-step baselines. The following visual examples demonstrate FlashEdit's ability to achieve a superior balance of fidelity and precision, consistently

Figure 3: More comparative visual results.

performing localized edits with high semantic accuracy while maintaining perfect background integrity. These results visually reinforce the quantitative findings of our main paper, confirming that FlashEdit is a robust and highly effective solution for real-time generative editing.

REFERENCES

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *ICLR*, 2023.

Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *ECCV*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801, 2008. doi: 10.1049/el:20080522. URL `https://digital-library.theiet.org/doi/abs/10.1049/el%3A20080522`.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, et al. A versatile diffusion transformer with mixture of noise levels for audiovisual generation. *arXiv preprint arXiv:2405.13762*, 2024.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. In *NeurIPS*, 2021.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Weizhu Chen, and Nan Duan. Genie: Large scale pre-training for text generation with diffusion model. *arXiv preprint arXiv:2212.11685*, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICCV*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Vidya Prasad, Chen Zhu-Tian, Anna Vilanova, Hanspeter Pfister, Nicola Pezzotti, and Hendrik Strobelt. Unraveling the temporal dynamics of the unet in diffusion models. *arXiv preprint arXiv:2312.14965*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *ArXiv*, abs/2202.00512, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Zhaohu Xing, Liang Wan, Huazhu Fu, Guang Yang, and Lei Zhu. Diff-unet: A diffusion embedded network for volumetric segmentation. In *MICCAI*, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.