Quality and Speed are All You Need

Youzhang He, Yiyan Liang, Yixiao Ling, Tongtong Liu
University of Pennsylvania
ESE 6450: Deep Generative Models (Fall 2024) - Final Project Report
{youzhang, edgarl, liing, liufrank}@seas.upenn.edu

Abstract

Text-to-image diffusion models have transformed the field of image synthesis, enabling users to generate diverse and realistic imagery based solely on textual prompts. However, transitioning from creative text-to-image generation to practical image editing remains challenging. While diffusion-based models offer remarkable fidelity and variety, controlling local edits without compromising global content is non-trivial. This paper critically examines four diffusion-based image editing approaches: DDIM, Null Text Inversion, InstructPix2Pix, and Pix2Pix Zero. Employing the PIE-Bench dataset—consisting of 700 images and a spectrum of 10 diverse editing categories—we systematically evaluate each method's ability to preserve structural fidelity, achieve semantic alignment with textual instructions, and balance computational costs. Additionally, we introduce and discuss consistency models that can significantly reduce inference times and offer inversion capabilities critical for interactive editing applications. Through extensive quantitative and qualitative analysis, our study highlights the trade-offs between structural preservation, semantic adherence, and computational efficiency. Our findings provide a comprehensive evaluation framework for guiding future research and development of user-driven, text-guided image editing techniques.

1. Introduction

The rapid advancement of generative modeling has shifted the landscape of image synthesis and editing. Text-to-image models, propelled by improvements in Generative Adversarial Networks (GANs) [4], Variational Autoencoders (VAEs) [10], and more recently, diffusion-based methods [6], have demonstrated unprecedented capabilities in creating detailed, photorealistic images from textual descriptions. Despite these successes, the practical application of such models for controlled image editing—where users refine a given image according to specific textual instructions—remains an unsolved challenge. Achieving fine-grained edits that maintain the original image's global structure, local details, and overall quality is notably difficult.

Existing text-to-image frameworks often suffer from overreliance on global conditioning, making localized adjustments unpredictable and sometimes destructive to unrelated parts of the image. To address these shortcomings, researchers have explored targeted interventions, such as masked generation, image-to-image translation, and text-guided latent manipulations.

In this paper, we offer a systematic and comprehensive evaluation of four baseline diffusion-based image editing methods: DDIM [23], Null Text Inversion [13], InstructPix2Pix [2], and Pix2Pix Zero [16]. To ensure rigor and reproducibility, we leverage the PIE-Bench dataset [9]—a curated collection of 700 images and 10 editing categories encompassing a wide range of real-world editing tasks, from object insertion and deletion to fine-grained attribute modifications and style changes. Using perceptual (LPIPS), structural (SSIM), and semantic (CLIP Similarity) metrics, we delve deep into each method's strengths and weaknesses. Additionally, we investigate how newer paradigms, such as consistency models and invertible consistency distillation (iCD) [25], promise substantial speed-ups and improved control, potentially paving the way for interactive editing workflows.

Our contributions are twofold: (1) We provide a holistic comparative study of prominent diffusion-based editing techniques, offering insights into the trade-offs between computational efficiency, structural fidelity, and semantic alignment. (2) We discuss the potential of consistency models and related innovations that could further streamline and enhance the editing process.

The GitHub repository for this project is available at: link

2. Related Work

In this section, we introduce text-to-image models in 2.1, followed by image editing with diffusion models in 2.2. Finally, we present the four baseline models evaluated and compared in this project in 2.3 and consistency distillation

method in 2.4.

2.1. Text-to-Image Models

Text-to-image models aim to generate realistic and semantically meaningful images from textual descriptions. These models have rapidly evolved due to advancements in deep learning, particularly with generative models such as Generative Adversarial Networks (GANs) [4], Variational Autoencoders (VAEs) [10], and diffusion models [6]. However, despite their success, these models typically lack fine-grained control over the generation process beyond the provided textual input. Real-world image editing remains challenging with these models, as altering the input text often leads to unexpected changes across the entire image.

Some approaches address these limitations by using masks to constrain where edits are applied [1, 14]. Others focus on training conditional diffusion models specifically designed for image-to-image translation tasks [2, 22, 26]. These methods enhance the control and precision of image modifications, bridging the gap between text-to-image generation and practical image editing.

2.2. Diffusion-Based Image Editing

Diffusion models [6] have emerged as powerful tools in generative modeling, capable of producing high-quality and realistic images. Beyond image generation, these models have demonstrated significant potential in image editing, enabling targeted modifications to existing images based on user-provided instructions or constraints.

SDEdit [12] showcased the ability of diffusion models to edit images by reintroducing noise into an existing image and guiding the denoising process using new conditions, such as text or sketches. OpenAI's GLIDE [14] further extended diffusion models' capabilities by combining text conditioning with guided diffusion, allowing users to provide text prompts for fine-grained image modifications. Another notable model, DALL-E 2 [18], introduced inpainting capabilities, where specific regions of an image could be modified while preserving the surrounding context. This was achieved by masking parts of the input image and applying diffusion-guided generation to the masked areas.

We focus on four well-established models [2, 13, 16, 23] for image editing tasks, which will be discussed in detail in the following section.

2.3. Baseline Models for this project

2.3.1. DDIM+Stable diffusion

DDIM [23] is a method for efficient image generation based on diffusion models, a class of generative models that learn to generate data by gradually denoising a sample starting from pure noise. While standard diffusion models rely on thousands of denoising steps to ensure quality, DDIM significantly reduces this number by using a non-Markovian sampling process. It achieves this through a reformulation of the original diffusion process, allowing for faster and more deterministic sampling while preserving the quality of generated outputs.

The key innovation in DDIM is the introduction of a parameterized sampling scheme that can traverse the generative process in fewer steps. By leveraging deterministic paths in the latent space, DDIM eliminates the stochastic noise component in sampling, making it particularly useful in applications where reproducibility and computational efficiency are critical. This makes it a foundational approach in modern generative frameworks, especially for large-scale, high-quality image generation.

Stable Diffusion [19] is a deep learning-based generative model designed for high-quality image synthesis, with a focus on text-to-image generation. It is based on diffusion models, where the model learns to reverse a diffusion process that progressively adds noise to training images. Unlike earlier diffusion models, Stable Diffusion introduces an efficient architecture leveraging latent space diffusion. Instead of operating directly on high-dimensional pixel spaces, it operates on a compressed latent space derived from a pre-trained variational autoencoder (VAE). This drastically reduces computational requirements while retaining visual fidelity.

We combine DDIM and stable diffusion together. DDIM provides an efficient sampling method for diffusion models, while Stable Diffusion leverages a diffusion process to generate images in a compressed latent space. We use stable-diffusion-v1-5 for the experiment.

2.3.2. Pix2Pix Zero

Pix2Pix Zero [8] is a novel framework for zero-shot image-to-image translation that utilizes pretrained diffusion models without requiring paired datasets or task-specific fine-tuning. Traditional methods like Pix2Pix [8] rely on supervised learning with paired data to learn mappings between source and target images, which limits their generalizability. Pix2Pix Zero addresses this limitation by leveraging pretrained text-to-image diffusion models (e.g., Stable Diffusion [19]) to guide image edits using natural language prompts. Pix2Pix Zero uses DDIM [23] for inversion. DDIM enables deterministic and efficient mapping of real images into the latent space of the diffusion model, forming the foundation for semantic edits driven by user-provided prompts.

By combining DDIM-based inversion with the generative capabilities of pretrained diffusion models, Pix2Pix Zero achieves zero-shot performance across a wide variety of tasks, including attribute editing, style transfer, and object replacement. While Pix2Pix Zero demonstrates significant generalization and versatility, its performance is influenced by the limitations of the underlying pretrained diffusion model, including biases in the training data and the

quality of the latent representation. Additionally, it uses other large language models like GPT-3 [3] to generate captions and sentences based on source and target prompt to learn the directional embeddings, which makes the inference very slow.

2.3.3. Null Text Inversion

Null Text Inversion [13] introduces a novel optimization strategy for classifier-free guidance diffusion models. Unlike prior methods that focus on optimizing model weights or text embeddings, this approach specifically targets the optimization of null-text embeddings used in unconditional generation.

In practical applications, DDIM inversion can introduce errors at each generation step. While such errors may be negligible for unconditional diffusion models, they can accumulate in classifier-free guidance diffusion models, causing the generated noise vector to deviate from the Gaussian distribution. When these vectors undergo DDIM sampling, the resulting images may deviate significantly from the originals, potentially introducing visual artifacts.

Traditional approaches [20] to address these issues often rely on either fine-tuning network weights, which risks compromising the valuable prior knowledge of the model, or optimizing text embeddings, which may result in semantically ambiguous or less interpretable outcomes. In contrast, null-text inversion offers an elegant and effective solution by focusing on the optimization of null-text embeddings.

This method provides several key advantages:

1. It achieves highly realistic editing results while preserving the semantic integrity of the editing operations.
2. It retains the model's pre-trained knowledge by avoiding direct modifications to network weights.
3. It enhances interpretability by maintaining clear semantic relationships within the text embeddings.

2.3.4. InstructPix2Pix

InstructPix2Pix presents a novel framework for instruction-guided image editing, leveraging natural language prompts to direct modifications to existing images. This model builds on advancements in text-to-image synthesis, such as DALL·E [18] and CLIP [17], which align textual and visual representations in shared latent spaces, and diffusion-based methods like Stable Diffusion [19], which optimize image generation processes. By fine-tuning a latent diffusion model on paired image edits and text instructions, InstructPix2Pix achieves diverse, context-aware image editing capabilities. The framework's ability to map free-form textual instructions to precise pixel-level modifications distinguishes it from task-specific methods like Palette and SDEdit, which often rely on structured constraints.

One of the primary advantages of InstructPix2Pix is its generalizability, making it applicable to a wide range of image editing tasks, from creative transformations to professional applications. The integration of instruction-following methodologies, inspired by NLP models like InstructGPT [15], ensures that edits align closely with user intent. However, the approach is not without limitations. The model's reliance on high-quality paired datasets may restrict its performance on out-of-domain instructions or unconventional editing tasks. Despite these limitations, InstructPix2Pix represents a significant step forward in interactive, multimodal AI systems, providing a robust foundation for future advancements in instruction-driven visual editing.

2.4. Consistency Models

Different from diffusion based models, consistency models[24] aim to directly learn a *consistency function* $f : (x_t, t) \longrightarrow x_\epsilon$, which satisfies the property of *self-consistency*: its outputs are consistent for arbitrary pairs of (x_t, t) that belong to the same *Probability Flow* (PF) ODE trajectory, which takes the form of

$$dx_t = \left[\mu(x_t, t) - \frac{1}{2}\sigma^2(t)\nabla_{x_t} \log p(x_t)\right] dt$$

Consistency functions could facilitate the image generation process in very few steps. Although consistency models could be trained from sketches, distillation from pre-trained models is a better choice as it skips the long training period. For example, Latent Consistency Models[11] (LCM) provide a practical way to distill, from pre-trained stable diffusion (SD), a latent denoising diffusion model.

Since traditional consistency models only contains a consistency function, so it can hardly do inversion as what DDIM or other denoising models could do. To apply consistency models to image-editing tasks, it should be prioritized to implement an inversion-like function inside the consistency models, which could clip both the latent visual representation as well as the prompt tokens, so that the conditional guidanced image edition could be possible. To address such a problem, inversible Consistency Distillation[25] (iCD) introduced a new way to distill from a pre-trained SD model by doing consistency distillation on both forward and reverse process of the pre-trained denoising models. Besides, they introduced a multi-boundary consistency distillation to make sure that iCD could do a similar multi-step consistency sampling as regular consistency models. Also, since the rate of classifier-free guidance[5] (CFG) plays a great role in the process of generation, a dynamic guidance[21] scheme CADS, was implemented to facilitate the inversion.

With the aid of iCD, image-editing task could be done in simply 2 phases. In the first phase, iCD extract the latent representation of the images via inversion, and clip it to the caption tokens. In the second phase, iCD do the conditional guidanced generation simply as traditional con-

sistency models. For both phases, 4-8 steps are enough to generate high quality images, which saves way much time compared to 30-50 steps that DDIM need.

3. Methodology

In this section, we discuss how we evaluate baseline models on PIE-Bench [9], as detailed in section 3.1. Additionally we specify how we evaluate the performance of consistency models in section 3.2.

3.1. Evaluating Baseline Models on PIE-Bench

We evaluate four baseline methods for diffusion-based image editing tasks: DDIM [23], Null Text Inversion [13], InstructPix2Pix [2], and Pix2Pix Zero [16]. These methods are benchmarked using PIE-Bench [9], a comprehensive dataset specifically tailored for diffusion-based image editing evaluation. Each method differs in its architectural design and editing strategy, offering distinct advantages and trade-offs.

PIE-Bench consists of 700 images paired with editing prompts spanning 10 diverse editing categories, such as modifying object attributes, adding new objects, or removing specific elements. It includes a balanced variety of natural and synthetic images, ensuring that the evaluation captures the editing performance of models across a wide range of challenging scenarios.

To quantitatively assess the performance of these baseline models, we adopt the following three metrics:

1. **Structural Similarity Index Measure (SSIM)** $\uparrow$
   SSIM is a classical image quality metric that measures the structural similarity between the edited and the original images. It evaluates luminance, contrast, and structural details. *A higher SSIM score* (closer to 1) indicates better preservation of the original image's structural fidelity.
2. **Learned Perceptual Image Patch Similarity (LPIPS)** $\downarrow$
   LPIPS quantifies perceptual similarity by comparing intermediate features extracted from a pre-trained deep neural network. Unlike pixel-based metrics, LPIPS is sensitive to human perceptual judgments. *A lower LPIPS score* (closer to 0) suggests a closer alignment between the edited image and human visual expectations.
3. **CLIP Similarity (Target)** $\uparrow$
   CLIP Similarity measures the alignment between the edited image and the target textual prompt using CLIP's feature embeddings [17]. This metric evaluates the semantic coherence of the edited image relative to the editing instructions. *A higher CLIP score* indicates a better semantic match between the edited image content and the textual description.

We also measure the time it takes (in second) to generate one image for each model. We present the results and further discussion in Section 4.1.

3.2. Consistency Models

To improve the performance of the evaluation, we introduced LoRA[7] to the pre-trained iCD[25] forward and reverse model. Specifically, we utilize LoRA to fine-tune the pre-trained model on a subset of the Pie-Bench data, and evaluate on another disjoint subset of images ($\sim$ 420 items). We evaluate the model performance based on following metrics:

1. **CLIP Score** $\uparrow$
   CLIP Score is a metric used to evaluate the semantic consistency between text and images. It is based on the CLIP model developed by OpenAI. Through contrastive learning, the CLIP model maps images and text into a shared high-dimensional vector space, enabling the measurement of their similarity. It leverages the embedding capabilities of the CLIP model to compare the cosine similarity between the embedding vectors of images and text. A higher score indicates stronger semantic consistency between the image and the text.
2. **ImageReward Score** $\downarrow$
   ImageReward is a model or framework focused on evaluating the quality of image generation. Its core goal is to quantify how well a generated image aligns with a target description by leveraging both matching evaluation and image quality assessment, and thus provide a good evaluation to assess the performance of image generation systems.

4. Results and Discussion

In this section, we present and discuss the outcomes of our experiments on diffusion-based image editing. We first report evaluation results on PIE-Bench [9] for the baseline models in Section 4.1, followed by an exploratory analysis of Null-Text Inversion in Section 4.2. Next, we provide evaluation results and a demonstration of the DDIM + Stable Diffusion approach in Section 4.3. Finally, we discuss the performance of consistency models in Section 4.4.

4.1. Evaluation Results on PIE-Bench

Table 1 summarizes the performance of four baseline methods, reporting SSIM, LPIPS, CLIP Similarity, and inference time. The reported inference time (in seconds) is the duration required to generate one image using each model with 50 inference steps. Bold text indicates the best model, and red text indicates the worst model.

DDIM [23] provides a foundational sampling approach for image generation, modifying the sampling process to be non-Markovian and thus more efficient. While DDIM achieves moderate SSIM (0.69) and LPIPS (0.17), its relatively low CLIP Similarity (21.22) indicates difficulty in aligning the edited image with the semantic target prompt.

Baseline	SSIM ↑	LPIPS ↓	CLIP Similarity (target) ↑	Inference Time (s)
DDIM [23]	0.69	0.17	21.22	50
Null Text Inversion [13]	**0.80**	**0.05**	22.76	91
InstructPix2Pix [2]	0.34	0.67	**26.28**	**5**
Pix2Pix Zero [16]	0.67	0.20	26.04	32

Table 1. PIE-Bench Evaluation Metrics and Inference Time Comparison

This limitation arises because DDIM itself is a sampling strategy rather than a text-conditioned generation framework.

Null-Text Inversion [13] enhances DDIM by incorporating an inversion-based optimization strategy. By identifying a null-text embedding that accurately reconstructs the input image, Null-Text Inversion achieves the highest SSIM (0.80) and lowest LPIPS (0.05) of all baselines, indicating strong structural and perceptual fidelity. However, its CLIP Similarity (22.76) is lower than the instruction-following baselines, suggesting that while it preserves image quality, it does not strongly emphasize semantic alignment with the target prompt.

InstructPix2Pix [2] represents a shift in methodology, explicitly training on text-based image editing instructions. It excels in semantic alignment (CLIP Similarity = 26.28), significantly outperforming the other baselines in this regard. However, this comes at the cost of structural and perceptual fidelity (SSIM = 0.34, LPIPS = 0.67). In other words, it achieves excellent adherence to the prompt at the expense of preserving fine-grained image details. Notably, InstructPix2Pix is also the fastest method tested (5 s), making it practical for rapid prototyping where prompt fidelity is more important than exact image preservation.

Pix2Pix Zero [16] balances these trade-offs by combining inversion techniques and zero-shot image-to-image translation. It achieves moderate to good results across all metrics (SSIM = 0.67, LPIPS = 0.20, CLIP Similarity = 26.04) and has a moderate inference time (32 s). This balanced performance suggests that Pix2Pix Zero can serve as a versatile option when both quality and prompt adherence are important considerations.

In summary, the results reveal a trade-off between structural fidelity, semantic alignment, and computational efficiency. Null-Text Inversion excels at preserving image quality but is computationally expensive and less semantically aligned. InstructPix2Pix is fast and semantically faithful but sacrifices image fidelity. Pix2Pix Zero strikes a balance between these extremes, while DDIM provides a foundational approach with moderate performance and speed. The choice of model should depend on the application requirements—high-fidelity but slower generation, prompt fidelity with rapid output, or a balanced compromise.

Figure 1. Qualitative comparison of diffusion-based image editing baselines. The left image is the input (a cat), and the editing prompt is "change the cat to a horse." The middle image is generated by Null-Text Inversion, and the right image is generated by InstructPix2Pix.

Figure 1 qualitatively illustrates these trade-offs. Null-Text Inversion preserves structural details such as the background, mirror reflections, and wood texture. This matches its quantitative performance, where it excels in metrics reflecting structural fidelity. InstructPix2Pix, while successfully following the prompt to replace the cat with a horse, introduces notable background changes and loses some textural fidelity, reflecting its emphasis on semantic alignment over structural preservation.

4.2. Exploratory Analysis on Null-Text Inversion

We conducted a series of exploratory experiments to understand how Null-Text Inversion behaves under various editing prompts. Figure 2 presents examples of these experiments.

First, we tested the model with more complex concepts in the prompt (second image). The model sometimes failed to represent all concepts (e.g., the "mirror" and reflective properties of the horse). This may be due to limited attention capacity and complex interactions in the latent space. In text-to-image models, the attention mechanism distributes finite resources across multiple prompt attributes. When prompts contain multiple or competing concepts, it becomes challenging for the model to isolate and balance these attributes. Additionally, training data bias can influence which concepts are more readily represented, favoring common attributes over rarer or more intricate ones.

Second, we examined how the model handles changes in focus within the editing prompt (third image). When modifying subtle attributes (e.g., "circle eyes" to "square eyes"), the model struggled to capture the differences. Dominant

Figure 2. Model behavior with varying editing prompts. The first image is the original input with the prompt: *"A cat sitting next to a mirror"*. The second image uses a complex prompt: *"A small horse with folded ears, wearing a blue collar, sits near a reflective mirror, its entire body mirrored in the glass backdrop."* The third image explores focus changes from *"A cat with circle eyes sitting next to the mirror"* to *"A cat with square eyes sitting next to the mirror"*. The fourth image evaluates lighting changes with the prompt: *"A cat sits beside a mirror under sunshine."*

Figure 3. Exploratory analysis on self-replacement steps.

features (such as "eyes" in general) might occupy more prominent latent space dimensions than subtle variations like shape differences. Since Null-Text Inversion adjusts embeddings in aggregate, these subtle distinctions can be overshadowed by dominant shared features.

Third, we introduced lighting changes (e.g., "sunshine"). The results indicated that incorporating lighting cues degraded overall generation quality. Null-Text Inversion focuses on learning static embeddings for tokens, and changes in global attributes like lighting are more challenging to anchor within a single static embedding. Such transformations often require global adjustments to color, contrast, and shadows, which are not directly tied to specific object attributes.

In a subsequent experiment, we investigated the effect of varying the number of self-replacement steps in Null-Text Inversion (from the set $\{0.2, 0.4, 0.5, 0.6, 0.8\}$). Representative results are shown in Figure 3. Fewer self-replacement steps can lead to underfitting, where the placeholder token does not capture the intended semantics. Conversely, too many steps can cause overfitting, resulting in overly rigid embeddings. Our findings suggest that 0.4 self-replacement steps strike the best balance, producing outputs that remain faithful to the prompt while retaining variability and creativity.

4.3. Evaluation Results & Demonstration: DDIM + Stable Diffusion

We evaluated various tasks—such as adding objects, changing attributes, and modifying global properties—to assess the performance of DDIM + Stable Diffusion. The results are shown in Table 2.

Local edits, including adding or deleting objects and changing object attributes (e.g., pose or color), performed best across the metrics. For instance, adding an object achieved the highest SSIM (0.722) and a relatively low LPIPS (0.158), indicating strong structural preservation and minimal perceptual distortion. Tasks involving attribute changes in position or color also showed low LPIPS values, highlighting their effectiveness in making localized changes with minimal impact on overall image quality.

The object deletion task performed reasonably well, with SSIM = 0.680 and moderate LPIPS = 0.179. The solid CLIP score (17) indicates that object deletion aligns with semantic expectations, likely due to the model's capability to inpaint backgrounds and fill in plausible details learned from its training.

Global edits, such as changing style or attribute material, were more challenging. These tasks require coherent transformations across the entire image, making it difficult for diffusion models to maintain structural fidelity. As a result, these tasks had lower SSIM, higher LPIPS, and weaker CLIP alignment.

Task	SSIM	LPIPS	CLIP (edit)
0_random_140	0.677	0.183	21
change_object_80	0.686	0.190	20
add_object_80	0.722	0.158	22
delete_object_80	0.680	0.179	17
change_attribute_content	0.699	0.157	21
change_attribute_pos	0.716	0.144	20
change_attribute_color	0.715	0.146	19
change_attribute_material	0.698	0.183	23
change_background	0.700	0.172	22
change_style	0.678	0.188	24

Table 2. Performance of different editing tasks using DDIM + Stable Diffusion, reported via SSIM, LPIPS, and CLIP scores.

Figure 4 shows a simple example where the prompt was "change the goldfish to a shark." The model successfully replaces the goldfish with a shark while preserving the overall scene structure, demonstrating its competence in localized object edits.

Overall, diffusion models excel at localized edits but face challenges with global transformations. While CLIP guidance helps ensure semantic relevance, it is less effective for extensive style or material changes. Improving performance

Figure 4. Replacing a goldfish with a shark using DDIM + Stable Diffusion.

on global edits may require alternative text encoders, more powerful CLIP models, or specialized training strategies.

4.4. Evaluation Results: Consistency Models

Table 3 presents the evaluation results of iCD [25], tested with four inversion steps and four guided generation steps on a Mac M4 device. Despite being run on a Macbook M4, iCD outperforms the baseline DDIM + SD model. Additionally, the inference time decreased from 50 seconds (as shown in Table 1) to 39 seconds, with even faster performance expected on an A100 GPU. Notably, consistency models maintain comparable quality of generation performance.

Furthermore, consistency models exhibit strong capabilities in inversion and editing tasks on the PIE-Bench dataset. As illustrated in Figure 5, these models can generate high-quality inversions with minimal degradation in image quality, particularly for simpler scenes featuring single objects against simple backgrounds.

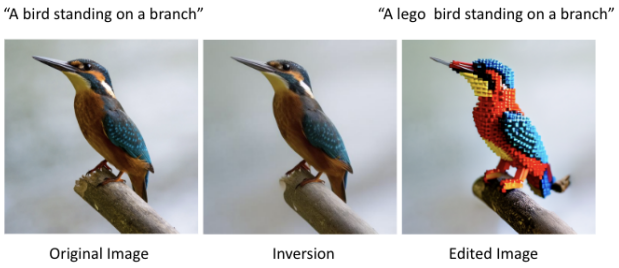

Figure 5. PIE-Bench image editing. The consistency model performs well on simple scenes with a single object and a uniform background.

However, as illustrated in Figure 6, more complex images with intricate backgrounds and multiple objects pose challenges. The inversion image may appear blurred, and the edited version can alter the background or distort object posture. This suggests that while consistency models hold promise, further research is needed to improve their robustness for complex, real-world scenes.

Figure 6. Real-world image editing with the consistency model. Complex backgrounds and multiple objects introduce artifacts such as blurred inversions and distorted object postures.

5. Conclusion and Future Work

In this study, we compared multiple baseline models on the PIE benchmark, revealing a fundamental trade-off between structural fidelity and semantic alignment. Each model demonstrated strengths in specific areas: DDIM served as a baseline for inversion, Null-Text Inversion excelled in structural preservation, InstructPix2Pix prioritized semantic alignment, and Pix2Pix Zero achieved a balance between these two aspects.

Our analysis highlighted several challenges associated with image editing models, specifically the Null-Text Inversion model, including limited attention allocation, error accumulation during generation, and difficulties in managing focus and lighting changes. Future research is needed to validate these findings across diverse tasks and explore solutions to enhance the model's robustness in addressing such challenges.

To tackle efficiency bottlenecks, we implemented consistency distillation. While invertible consistency distillation (iCD) demonstrated potential in accelerating the process and supporting inversion by maintaining forward and backward consistency through the use of a LoRA adapter, it remains limited in effectively handling complex scenes. Further development is required to improve its scalability and performance in diverse and challenging scenarios. Moreover, utilizing advanced text encoders will enable the model to better fetch the text information; and adding new visual encoders will help the model generate better latent space to benefit image editing. Also, ControlNet [27] could also be a good option to adopt so that the editing could be more precise and accurate, without distortion of background.

6. Contribution

Note: The order of authors is alphabetical and does not reflect the extent of individual contributions. We agreed that every member contributes roughly 25% to the project.

1. Tongtong is responsible for writing several parts of the report and implement the code related to two of the base-

CLIP Score (Inversion)	CLIP Score (Editing)	ImageReward Score	Inference Time (s)
0.8703	0.2529	0.04	39

Table 3. Performance of Consistency Models (iCD [25]).

line models and baseline models evaluation, including 1, 2.1, 2.2, 2.3.2, 2.3.4, 3.1, and 4.1. Tongtong coordinates the group for meetings and sets milestones.

2. Yixiao is responsible for writing the parts related to null text inversion, including 2.3.3, 4.2, and 5.
3. Youzhang is responsible for writing the parts related to DDIM+stable diffusion and build the corresponding code, including 2.3.1, 4.3, 5.
4. Yiyan is responsible for writing the parts related to consistency models and LoRA fine-tuning, and run the corresponding experiments, including 2.4, 3.2, 4.4, 5.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18187–18197. IEEE, 2022.

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

[9] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code, 2023.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[11] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

[12] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022.

[13] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[15] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[16] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[18] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[21] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling, 2024.

[22] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[24] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

[25] Nikita Starodubcev, Mikhail Khoroshikh, Artem Babenko, and Dmitry Baranchuk. Invertible consistency distillation for text-guided image editing in around 7 steps. *arXiv preprint arXiv:2406.14539*, 2024.

[26] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation, 2022.

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.