Quality and Speed are All You Need

ESE 6450: Deep Generative Models (Fall 2024) - Final Project

Youzhang He, Yiyan Liang, Yixiao Ling, Tongtong Liu

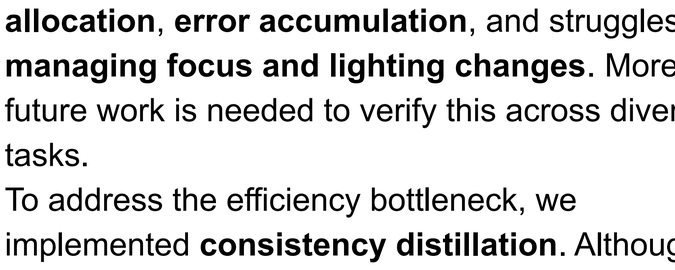
{youzhang, edgarl, liing, liufrank}@seas.upenn.edu

0. Introduction

- Diffusion model based image editing is an interesting but challenging tasks, especially on real images
- In this project, we evaluate and compare the performance of four baseline models in PIE-Bench using three different metrics
- We explore capacity of various models on different image-editing tasks
- We implement consistency distillation technique to improve the inference speed; it only takes 4-8 steps to generate or edit images

2. Stable diffusion+DDIM

- **Efficient Text-to-Image Generation with Stable Diffusion**: Combines latent space diffusion with high-resolution support to produce realistic and customizable images from text prompts, widely used in creative industries.
- **Accelerated Sampling with DDIM**: Reduces the number of steps required for image generation, offering faster inference while maintaining high image quality and smooth transitions.
- **Enhanced Control and Quality**: Integrates DDIM's non-Markovian sampling with Stable Diffusion, enabling precise control over the generation process for diverse applications
- **Pros and Cons**: Good at making local modifications, but still has room of improving on style transferring and global editing such as material editings
- **Future Improvement**: Use style CLIP as text encoder, and add additional vision encoder to better encode the details of the input images; or use Lora to make fine-tune on text encoders.



Prompt: Change the goldfish in the picture to shark

6. Conclusion

- We compare multiple baseline models on PIE-Benchmark. Our findings highlights the **trade-off between structural fidelity & semantic alignment**. Each model specialize in specific areas: DDIM provides baseline inversion, Null-Text Inversion succeeds in **structural preservation**, InstructPix2Pix prioritizes **semantic alignment**, and Pix2Pix Zero **balancing** both aspects.
- We then explored the challenges faced by null-text inversion model, such as **limited attention allocation**, **error accumulation**, and struggles with **managing focus and lighting changes**. More future work is needed to verify this across diverse tasks.
- To address the efficiency bottleneck, we implemented **consistency distillation**. Although invertible consistency distillation (iCD) shows some power in accelerating the process while supporting inversion by **ensuring forward and backward consistency with Lora adapter**, it still faces challenges in handling **complex scenes**.

1.Baseline Evaluation Results on PIE-Bench

Baseline	PIE-Bench Evaluation Metrics			Inference Time for One Image (seconds)
	SSIM ↑	LPIPS↓	CLIP Similarity(target) ↑	
DDIM	0.69	0.17	21.22	50
Null Text Inversion	**0.80**	**0.05**	22.76	91
InstructPix2Pix	0.34	0.67	**26.28**	**5**
Pix2Pix Zero	0.67	0.20	26.04	32

- **DDIM**: provides a baseline for inversion but lack fine control for semantic alignment with the lowest CLIP Similarity score
- **Null-Text Inversion**: best structural fidelity due to precise inversion with the cost of sacrificing semantic alignment with highest SSIM and LPIPS score
- **InstructPix2Pix**: excels in semantic alignment due to instruction-following training but poor structural preservation with highest CLIP similarity
- **Pix2Pix Zero**: balanced structural fidelity and semantic alignment. Slightly lower performance comparing to null-text inversion due to using DDIM inversion
- **Takeaways**:
  - trade-off between structural fidelity (SSIM and LPIPS) and semantic alignment (CLIP Similarity)
  - Accurate inversion techniques is important for preserving structural fidelity
  - Different baseline models focuses on various tasks

Input Image Null-text Inversion InstructionPix2Pix

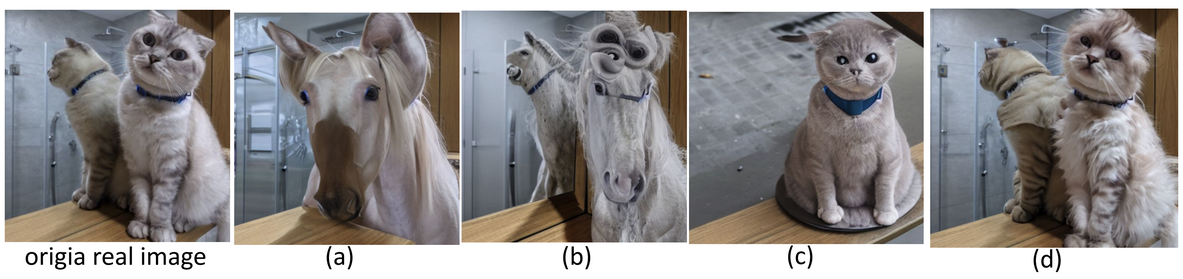
3. Exploratory Experiments

(a) For the null text inversion model on object changing tasks, if more complex concepts are introduced in a prompt, the model may ignore some of them. **[limited attention allocation, semantic competition, training data bias]**

(b) More self replace steps increase image complexity, but may make images worse. **[error accumulation, overfitting, semantic drift]**

(c) The model struggles to manage focus changes:
"A cat with circle eyes sitting next to the mirror"
"A cat with square eyes sitting next to the mirror"

(d) The model struggles to capture light changes
"A cat with folded ears, wearing a blue collar, sits beside a mirror under sunshine."

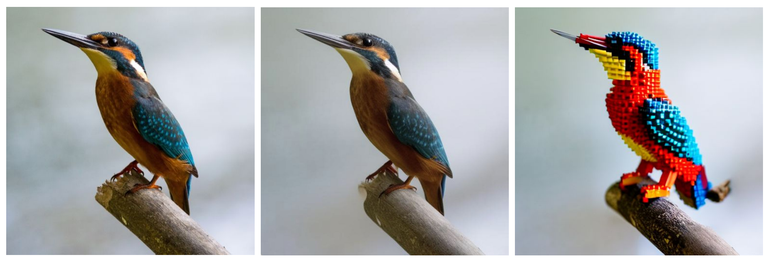


origia real image (a) (b) (c) (d)

4. Consistency Distillation

- Though **DDIM** significantly decreases the sampling steps, iit still take 30-50 steps to generate high quality images
- **Consistency Models** address this problem by learning a **consistency function** $f : (\mathbf{x}_t, t) \mapsto \mathbf{x}_\epsilon$ from pretrained DDIM models.The outputs are consistent for $(\mathbf{x}_t, t)$ that belong to the same **Probabilistic Flow** (PF) ODE trajectory
- However, consistency models don't support inversion, which is crucial for image-editing tasks. An **invertible consistency distillation** (**iCD**) was introduced to realize the inversion and distillation at the same time.
- **iCD** distills both forward and reverse process with **LoRA** adapter, and utilizes dynamic **CFG** guidance rate to better map the intermediate steps onto the PF ODE trajectories; **iCD** performs well on simple scenes (*left*), but for complicated scenes (*right*) this distillation could probably distort the inversion and guidanced generation

"A bird standing on a branch" "A lego bird standing on a branch"



Original Image Inversion Edited Image

"A korean general rides white horse in snowfield" "An american general rides white horse in snowfield"

Original Image Inversion Edited Image

5. Comparison

- We compare the baseline models and the consistency distillation model, and we find that there is a trade off between the quality of generated images and the the inference time. Although null text inversion performs best, it also takes the longest inference time (roughly 110 seconds); consistency distillation model has a better balance between the inference time and image quality.
- Different models have different advantages and disadvantages. Moreover, the prompts and the sampling steps also play an important role for the image quality and the inference time. This is a hyperparameter tuning problem that should be tuned based on specific context and requirements.