Consistency Training Inverse: An Alignment Learning Approach for Representation Learning

Chia-Hong HSU
Department of Computer Science
Brown University
Providence, RI 02912
chia_hong_hsu@brown.edu

Abstract

Consistency Models generate high-quality images with few number of function evaluations (NFE) through consistency distillation or consistency training [13]. We introduce ***CT Inversion***, a novel approach that solves the initial value problem of the forward probability flow ODE from image distribution to Gaussian noise deterministically. Training with CT Inversion, our noiser model is able to encode the latent noise of a given image with 1 NFE as oppose to the baseline DDIM Inversion [10] that typically uses 35$\sim$100 NFE's. In addition to the acceleration on inference speed, our noiser model demonstrates close performance compared to DDIM Inversion in image reconstruction when coupled with off-the-shelf diffusion models or consistency models on the CIFAR-10 dataset. With classifier-free guidance, we further showcase on the MNIST dataset that CT Inversion can intuitively substitute DDIM Inversion in controlable generation algorithms such as image editing. In the end, we discuss future work on controllable generation by training CT Inversion on large scale image/text-image datasets. We also highlight its unique *alignment learning* training fashion as a new representation learning approach in comparison with traditional self-supervised methods using autoencoders or VAE's.

1 Introduction

Recent advancements in diffusion models (DM) have led to significant progress in synthesis of high-quality images. Consistency Models (CM) are a novel family of models designed to expedite the generation process by mapping noise to data through consistency distillation (CD) and consistency training (CT)[13]. CD samples images that lie on the same probability-flow (PF) ODE trajectory computed by a parent DM, then establishes consistency matching between adjacent images of different noise scales. However, CM's multi-step sampling procedure results in semantic variation as it involves alternating between zero noise and noise. Consistency Trajectory Models solves this problem by flexibly traversing backwards to any point on the PF ODE, with the option to jump forward for diverse generation controlled by a γ factor. [6]. Latent Consistency Models leverages consistency matching within the latent space of pretrained Stable Diffusion [12] autoencoders. Their guided CD method enables classifier-free guidance for few-step 2$\sim$4 or even 1-step sampling.

DMs and CMs, often referred to as denoisers, are recognized for their ability to transform noisy images into clean ones. The endeavor to solve the inverse problem, i.e., deducing the unique noise from a given clean image, has gained attention in the realm of inversion-based image editing works [10, 9, 3]. Building on the assumption that the ODE process can be reversed and approximated when considering infinitesimally small steps, DDIM Inversion performs diffusion in the reverse direction to find an image's noise encoding [10]. However, the inverse process is time-intensive due to the high number of function evaluations (NFE) of diffusion models.

Preprint. Under review.

We introduce Consistency Training (CT) Inversion, a novel approach for solving the initial value problem of the PF ODE from image distribution to Gaussian noise deterministically. Our method significantly reduces the NFE's required for noise encoding. It also achieves similar performance to the baseline DDIM Inversion on image reconstruction. In summary,

- We propose CT Inversion for deterministic noise encoding with only 1 NFE.
- We showcase that CT Inversion can be paired with off-the-shelf denoiser for image reconstruction tasks on CIFAR-10 [8] and MNIST [2]. With off-the-shelf DM's and CM's on CIFAR-10, CT Inversion provides 30 times speed up while having acceptable performance compared to the DDIM Inversion baseline.
- We also demonstrate its potential for controllable generation tasks via classifier-free guidance on the MNIST dataset. However, we leave the actual prompt-to-prompt image editing task for future work by training CT Inversion on large scale text-image datasets.
- Inspired by CM's, we introduce a novel *alignment learning* training style for representation learning in isolation, distinct from traditional self-supervised methods that uses autoencoders or VAE's [7].

2 Background

2.1 Consistency Models

Consistency Models (CM) are built upon the framework of diffusion stochastic differential equations (SDE). At $t = 0$, a real image sampled from the dataset $x_0 \sim D$, the generalized form describing the noisify process is given by:

$$\mathrm{d}x_t = \mu(x_t, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}w_t \tag{1}$$

Solving the Fokker-Planck Equation [11] yields the probability-flow Ordinary Differential Equation (ODE) as $\mathrm{d}x_t = [\mu(x_t, t) - \frac{1}{2}\sigma(t)^2\nabla \log(p_t(x_t))]\,\mathrm{d}t$. The absence of the stochastic term $\mathrm{d}w_t$ makes the evolution of x_t deterministic. For instance, in the Variance Exploding (VE) SDE [14] family with $\mu = 0$ and $\sigma(t) = \sqrt{2t}$, the corresponding PF-ODE becomes:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -t\nabla \log(p_t(x_t)) = \frac{x_t - \mathbb{E}[x_0|x_t]}{t} \tag{2}$$

Given the initial condition at $t = i$, $x = x_i$, as well as an optional ODE solver as an *oracle*, CM's are trained to model the landscape of the initial value problem (IVP) at $t = 0$, $x = x_0$. Denote the full trajectory of an image $\{x_t\}_{t\in[\epsilon,T]}$, the consistency property [13] is expressed as $f_\theta(x_{t+1}, t + 1) = f_\theta(x_t, t)$, $\forall t \in [\varepsilon, T]$. Consequently, CM's loss function is defined as:

$$L(\theta) = \mathbb{E}[\lambda(t)d(f_\theta(x_{t+1}, t + 1), f_{\theta^-}(x_t, t))] \tag{3}$$

Inspired by EDM's [5], CM's are parameterized with skip connections, $f_\theta(x_t, t) = c_{\text{skip}}(t)x_t + c_{\text{out}}(t)\text{NN}_\theta(x_t, t)$. This allows boundary conditions at $t = \varepsilon$ to propagate effectively to higher noise levels by setting $c_{\text{skip}}(\varepsilon) = 1$, $c_{\text{out}}(\varepsilon) = 0$, and $c_{\text{skip}}(T) = 0$, $c_{\text{out}}(T) = 1$. Consequently, $f_\theta(x_\varepsilon, \varepsilon) = x_\varepsilon$ skips the model's inference when input is approximately x_0, and $f_\theta(x_T, T) = \text{NN}_\theta(x_t, t)$ requires full model inference when the noisiest possible image is provided.

2.2 DDIM Inversion

The VE SDE's deterministic denoising process follows the euler update rule by:

$$x_{t-1} = x_t + \mathrm{d}x_t,\ \mathrm{d}x_t = \frac{x_t - \mathbb{E}[x_0|x_t]}{t}\,\mathrm{d}t \tag{4}$$

Following the common technique proposed in DDIM Inversion [10], we can approximate $\mathbb{E}[x_0|x_t]$ with $\mathbb{E}[x_0|x_{t-1}]$. By rearranging terms in Eq.4, the DDIM Inversion update rule is as follows:

$$
\begin{aligned}
x_t &= \frac{t}{t-1} x_{t-1} - \frac{\mathbb{E}[x_0|x_{t-1}]}{t-1} \, \mathrm{d}t \\
&= x_{t-1} - \frac{x_{t-1} - \mathbb{E}[x_0|x_{t-1}]}{t-1} \, \mathrm{d}t
\end{aligned} \tag{5}
$$

In practice, the expected value is approximated by a trained DM denoiser $g_\phi(x_t, t) \sim \mathbb{E}[x_0|x_t]$.

2.3 Inversion-based Image Editing

Null-text Inversion [10] ensures the fidelity of the original image by fine-tuning the null-text embedding in guided DM's [4]. Substituting the null-text placeholder with source text, Negative-prompt Inversion [9] accelerates image editing inference by 25 times with minimal performance degradation. With the use of a proximal function and mutual self-attention control [1], ProxEdit [3] prevents overestimation of the guidance strength without tuning parameters.

In this technical report, we will demonstrate CT Inversion's use case on changing MNIST digits by different digit classes. We show that CT Inversion + off-the-shelf CM's successfully changes the appearance of the input digit to the target class, while preserving the style of the original input as much as possible.

3 Image to Noise: Consistency Training Inversion

We propose CT Inversion, an algorithm that trains a model to directly map the source distribution to Gaussian noise by solving an IVP of the defined ODE in Eq.2. Specifically, through consistency matching, CT Inversion computes the noise latent x_T given any noise scale image x_i, $i \in [\varepsilon, T]$. For details, refer to the training objective in Alg.1.

Algorithm 1 Consistency Training Inversion (CT Inversion)

Require: dataset D, initial model parameter θ, learning rate η, step schedule $N(\cdot)$, loss function $d(\cdot, \cdot)$, weight schedule $\lambda(\cdot)$, and scale distribution P

$\theta = \theta^-$ and $k = 0$
repeat
  Sample $x_0 \sim D$ and $n \sim P$
  Sample $z \sim N(0, 1)$
  $x_n = x_0 + t_n z$
  $x_{n+1}(x_n, x_0) = x_n + (t_{n+1} - t_n)\frac{(x_n - x_0)}{t_n}$
  $L(\theta, \theta^-) \leftarrow \lambda(t_n) d(f_{\theta^-}(x_{n+1}, t_{n+1}), f_\theta(x_n, t_n))$
  $\theta^- \leftarrow \text{stopgrad}(\theta)$
  $k \leftarrow k + 1$
until convergence

As oppose to CM's, the propagation of the boundary condition flows backwards from $t = T$ to $t = \varepsilon$. We adopt the same skip connection structure as CM's and define $c_{\text{skip}}(t) = 1$, $c_{\text{out}}(t) = T - t$ such that it naturally satisfies $f(x_T) = x_T$. In case of low noise t, emphasis on the prediction is scaled up linearly in contrast to the skip connection of x_t. In practice, the input image x_t is also scaled by $c_{\text{in}}(t) = 1/\sqrt{t^2 + \sigma_{data}^2}$, which we choose to omit for clarity. It can be shown that such parametrization yields unit variance between the input and target, a nice property to have for training neural networks. The derivation of each parametrization for $c_{\text{skip}}(t), c_{\text{out}}(t), c_{\text{in}}(t)$ follows the Appendix in [5].

We highlight that the ground truth of an image's noise representation is unknown and intractable under CT Inversion training. Unlike traditional VAE that utilizes an autoencoder pipeline for self-supervision, CT Inversion is able to train the noise encoder ***in isolation*** without the help of a denoiser or an oracle ODE Solver. This, however, is based on the assumption that we are able to approximate clean images x_0 with low noise images x_ε. Otherwise, the uniqueness of the encoding process will not hold given the source distribution are dirac deltas in theory.

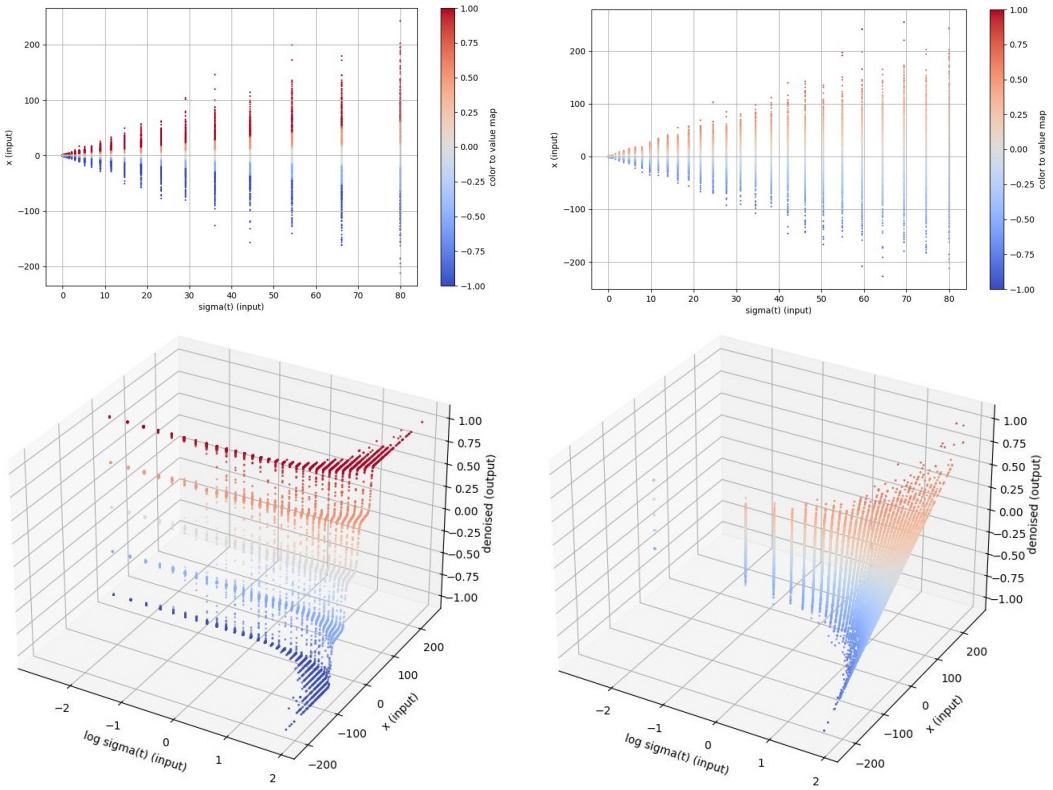

Figure 1: Toy illustration of CM denoiser (left) and CT Inversion encoder (right). The source data is uniformly distributed across five values in {-1.0,-0.5,0.0,0.5,1.0}. The xy-axis in the top row figures represents noise scale t and perturbed data x_t respectively. The color of each data point is the model's prediction, i.e., $f_\theta(x_t, t)$. BY lifting the value of each color up in 3d, we can see that CM learns to match the source data distribution given noisy data, while CT Inversion learns to match Guassian noise given clean data.

Corollary 3.1 *The stochastic process designed for consistency matching in CT Inversion matches the target density derived from Fokker-Plank Equation.*

Corollary 3.2 *The deduced target image x_{n+1} is dependent on x_0 and x_n, where x_n is obtained via the stochastic process that fulfills Corollary 3.1.*

Theorem 3.3 (Equivalence of CT Inversion's Objective and PF ODE Solution at time T) *Objective J_{CT} can be written as the following expected value:*

$$
\begin{aligned}
J_{CT} &:= \mathbb{E}_{x_0,n}[\lambda(t_n) d(f_{\theta^-}(x_{n+1}, t_{n+1}), f_\theta(x_n, t_n))] \\
&:= \mathbb{E}_{x_0,n}[\lambda(t_n) d(f_{\theta^-}(x_n + (t_{n+1} - t_n)\frac{(x_n - x_0)}{t_n}, t_{n+1}), f_\theta(x_n, t_n))] \\
&:= \mathbb{E}_{x_n,n}[\lambda(t_n) d(f_{\theta^-}(x_n + (t_{n+1} - t_n)\frac{(x_n - \mathbb{E}_{x_0}[x_0|x_n])}{t_n}, t_{n+1}), f_\theta(x_n, t_n))] \\
&:= \mathbb{E}_{x_n,n}[\lambda(t_n) d(f_{\theta^-}(x_n + dx_n, t_{n+1}), f_\theta(x_n, t_n))]
\end{aligned}
$$

Because of Corollary 3.1, the density $p(x_n)$ matches the PF-ODE density at $t = n$. Following the deterministic path by $x_n + dx_n$, the density $p(x_{n+1})$ also matches the PF-ODE assuming infinitesimal step sizes. Further assume that the objective J_{CT} converges to 0, this induces that images on the same PF-ODE trajectory $\{x_i\}_{i\in[\varepsilon,T]}$ are mapped to the same vector $f_\theta(x_T, T) = x_T$.

The effect of consistency matching on the denoising and encoding process is depicted in Fig.1. We call this new unsupervised training style *alignment learning* for representation learning tasks, which

Figure 2: MNIST reconstruction qualitative results. Left: ground truth batch (input to the noiser). Right: reconstruction batch (prediction from the denoiser). The noiser and denoiser are trained separately.

effectively aligns the source distribution to Gaussian noise. Without knowing the representation ground truth of individual samples from the dataset, we demonstrate in Sec.A that our encoder is able to communicate with a denoiser by reconstructing images from the CT Inversion's latent.

4 Experiments

To evaluate CT Inversion's proof of concept, we consider two simple datasets, MNIST and CIFAR-10, For the CIFAR-10 datasets, we further compare our results with DDIM Inversion as a baseline in terms of its lpips/ssim/psnr metric scores, as well as the average inferencing speed per batch data through the entire testing dataset. The DM baseline model, as well as the CM baseline models are downloaded from the open source github repo at `https://github.com/openai/consistency_models`.

We unify the UNet architecture for both the noiser and denoiser. Following training curriculum suggested in [13], we adopt a noise schedule by setting $\rho = 7.0$, maximum noise scale 80.0, minumum noise scale 0.002. For MNIST, the noiser models are trained about 40 A5000 gpu hours, completing 100K training iterations. As for CIFAR-10, the noiser models are trained about 288 A5000 gpu hours (8 gpus, 1.5 days), completing 300k training iterations. For image reconstruction tasks, our model is trained without guidance. In other words, both noiser and denoiser takes an image and its noise scale as input. For the experiments in Sec.4.2, the models take an additional class embedding as input in order to encode/generate class-conditioned results.

4.1 Image Reconstruction

For the MNIST dataset, the noiser is trained with CT Inversion, while the denoiser is trained with CT. The results show that despite being trained in isolation, the noise encoder can *talk to* another separately trained denoiser by reconstructing faithful images same as the input. Pairing CT Inversion + CT, we achieve high metric scores with lpips=0.0619, ssim=0.9270, and psnr=21.6735 (Fig.2).

According to the update rules in Eq.5, we set both the DDIM Inversion and DM inference steps to 30. We show a total of 4 experiments as shown in Table1, *DDIM Inversion + DM*, *CT Inversion + CM*, *DDIM Inversion + CM* and *CT Inversion + DM*. Similar to the results from MNIST, the noiser trained via CT Inversion is able to reconstruct faithful images when paired with off-the-shelf models. Surprisingly, when paired with DM's, CT Inversion even obtained marginal performance boost with the cost of inference time from the denoising process. We also observed that the reconstruction quality of CM-based methods are clearer than DM ones. However, metric-wise, CT Inversion still gets

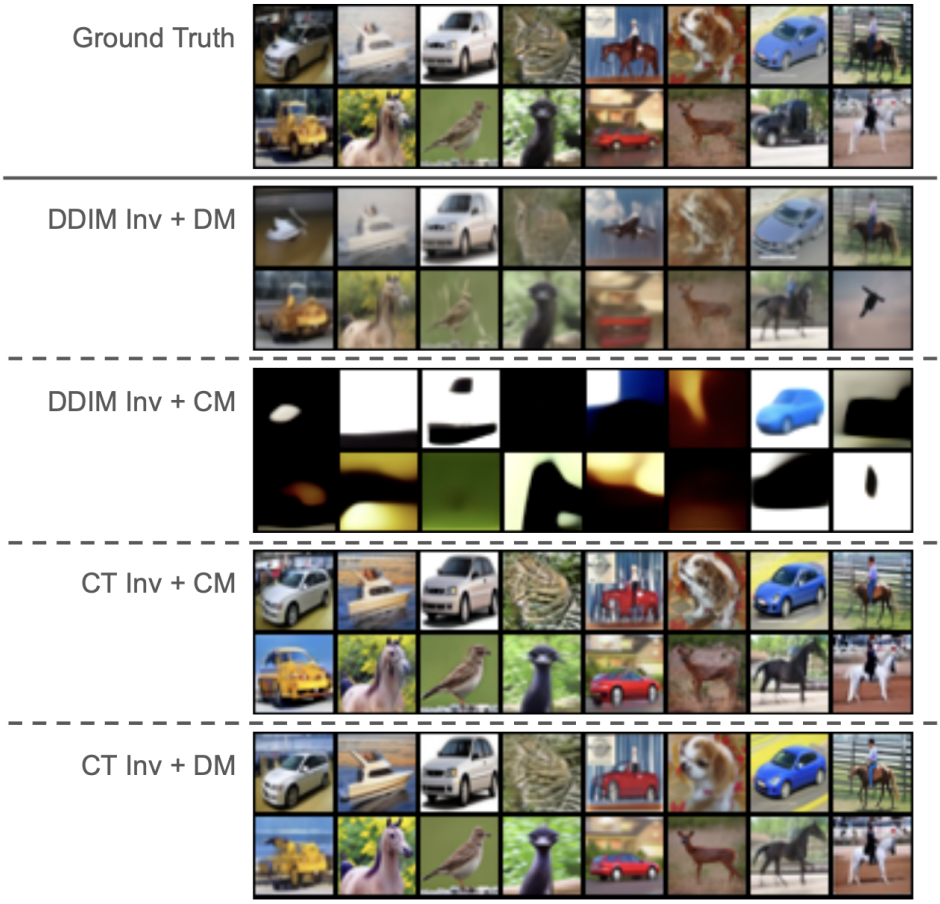

Figure 3: CIFAR-10 reconstruction qualitative results. With CT Inversion, the noiser is able to capture the denoiser's latent space. This is not possible for DDIM Inversion when pairing it with a CM denoiser. Overall, CT Inversion latents can be decoded into clearer images than their DDIM counterparts. However, noticeable tone deviation can be seen from the more contrasting images reconstructed by CT Inversion.

punished for the overall tone deviation from the ground truth. To conclude, CT Inversion achieves up to 40 times speed up with acceptable degradation on reconstruction quality. We present the qualitative results in Fig.3.

Table 1: CIFAR-10 reconstruction results

	lpips(↓)	ssim(↑)	psnr(↑)	cost (sec/batch)(↓)
DDIM_Inv + DM	**0.2165**	**0.8298**	**22.9758**	6.9768
DDIM_Inv + CM	0.6262	0.2499	10.3434	3.4931
CT_Inv + CM	0.2426	0.7911	18.0738	**0.1851**
CT_Inv + DM	0.2369	0.8024	18.1951	3.6525

4.2 MNIST: Style Preserving Digit Conversion

Inversion-based methods for controllable generation, e.g., image editing, are the main applications of DDIM Inversion for DM's. Previous works attempted to save the inference and optimization cost by eliminating the need of per-image fine-tuning. With CT Inversion, we can potentially further improve the inference cost significantly for fewer NFE's.

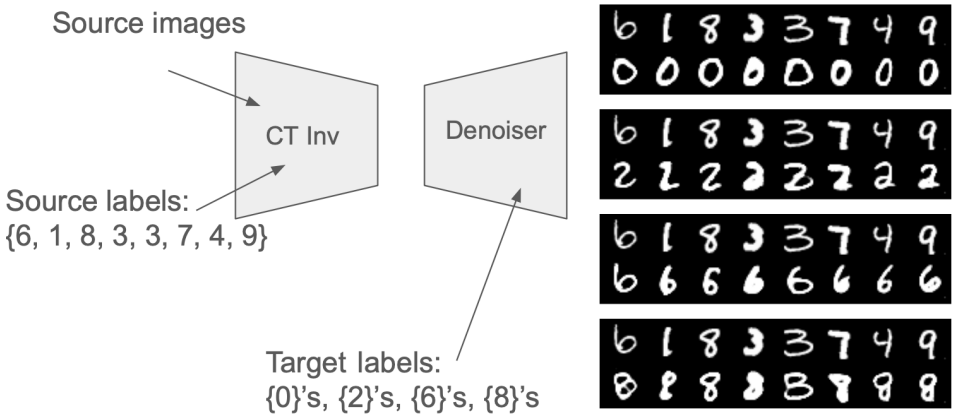

Figure 4: MNIST style preserving conversion qualitative results. The first row represents the source image, and the bottom row represents its conversion. CT Inversion manages to preserve the style of the source image even after conversion, e.g., stroke thickness, width length, thinness, source digit feature, etc.

In this section, we exhibit CT Inversion's capability of style preserving digit conversion on the MNIST dataset. Similar to the image editing task, we are given a source (image, class) pair, the objective is to modify the image according to a target class. As illustrated in Fig.4, CT Inversion is capable of converting the source image if the target class is different from the source, otherwise, it reconstructs the source image.

5 Closing

Inspired by consistency matching from CM's, CT Inversion reduces the computational burden associated with noise encoding. It's ability to achieve noise encoding with just 1 NFE surpasses baseline methods, ensuring rapid inference speed with minimal tradeoff in metric scores. With alignment training on a universal representation space, CT Inversion is capable of substituting DDIM Inversion in previous controllable generation algorithms without compatability concerns of the type of denoiser (as long as they are trained on the same source distribution).

Future work will extend CT Inversion to large-scale image and text-image datasets, maximizing its potential in inversion-based generation tasks like image editing. Furthermore, CT Inversion's alignment learning approach offers a unique path for representation learning, distinguishing it from traditional self-supervised methods. These efforts promise to advance generative models and enable more efficient and versatile applications across domains.

References

[1] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[2] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[3] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4291–4301, 2024.

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[5] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

[6] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

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

[8] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[9] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023.

[10] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[11] Hannes Risken and Hannes Risken. *Fokker-planck equation*. Springer, 1996.

[12] 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*, pages 10684–10695, 2022.

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

[14] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

A Appendix

Additional experimental results are shown here.

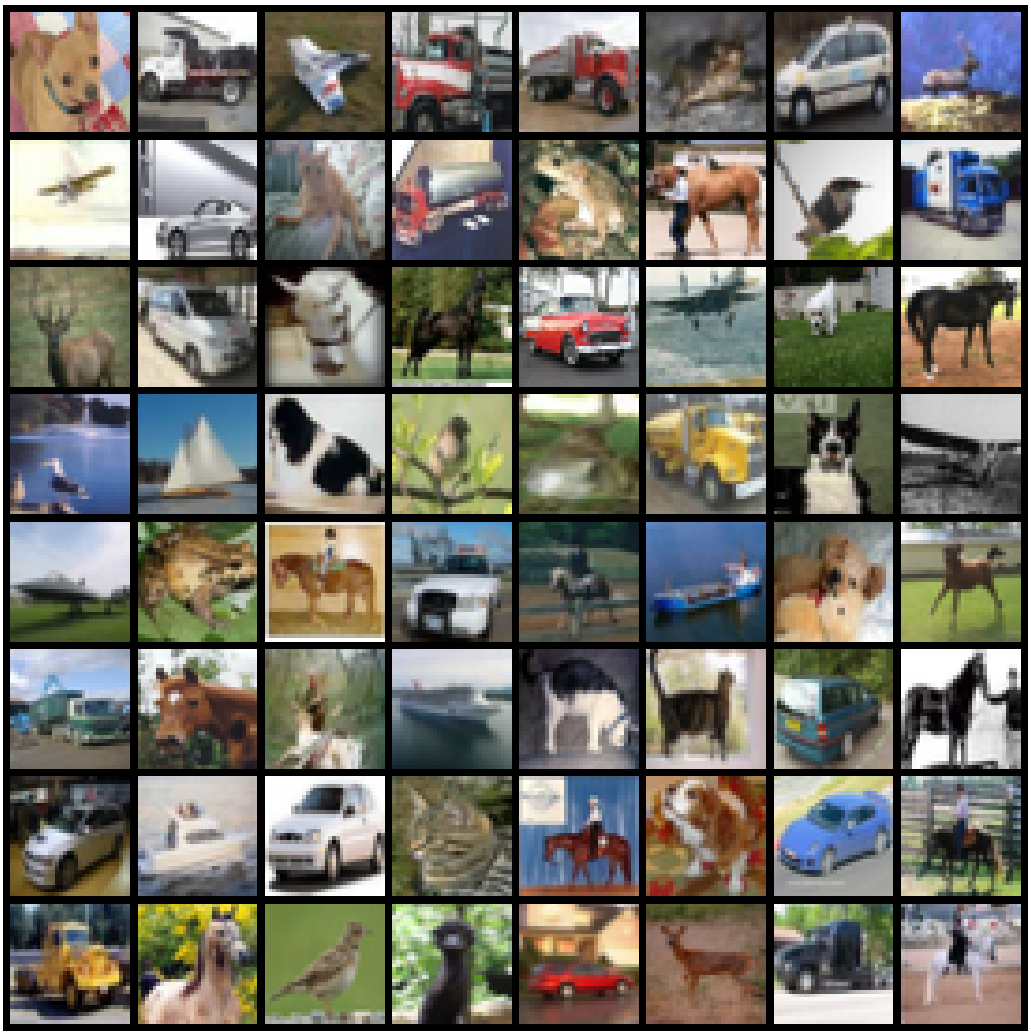

Figure 5: CIFAR-10 ground truth (single batch)

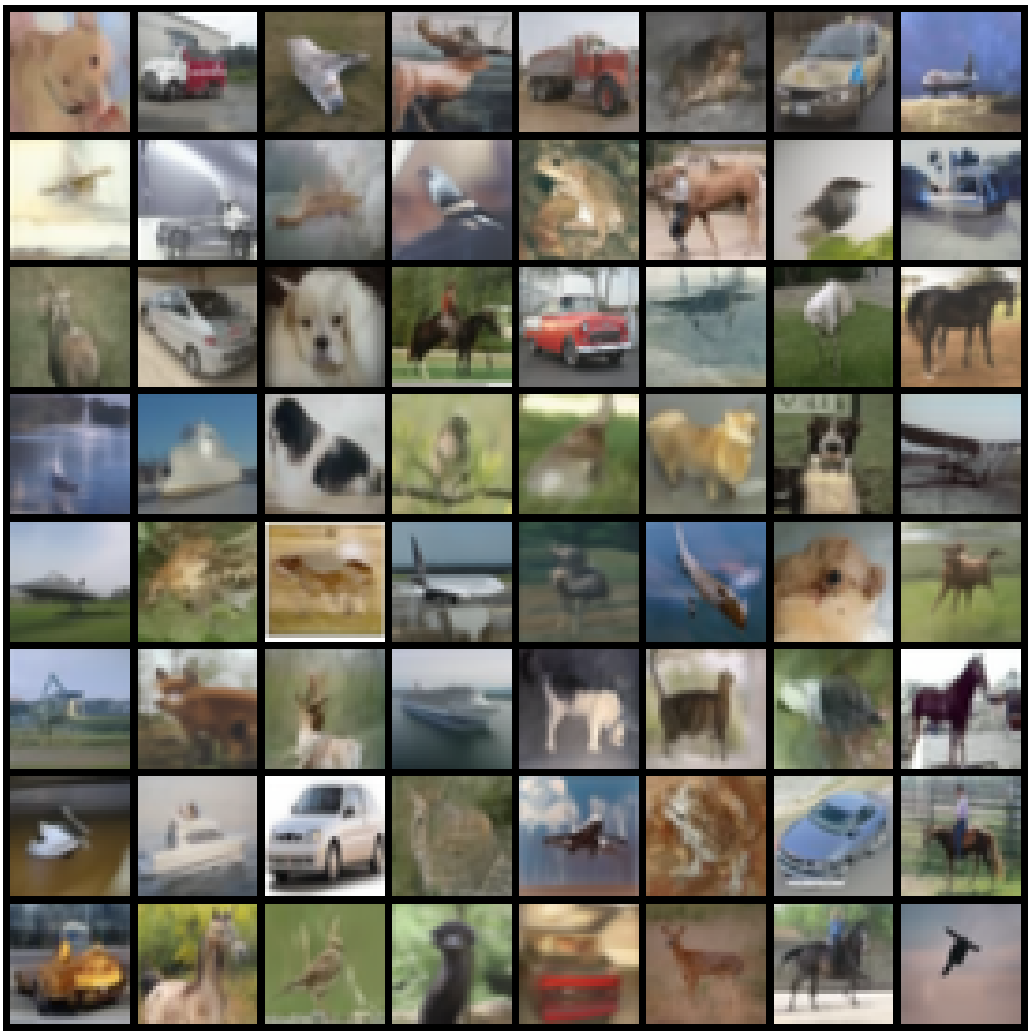

Figure 6: CIFAR-10 reconstruction: DDIM Inversion $\rightarrow$ DM

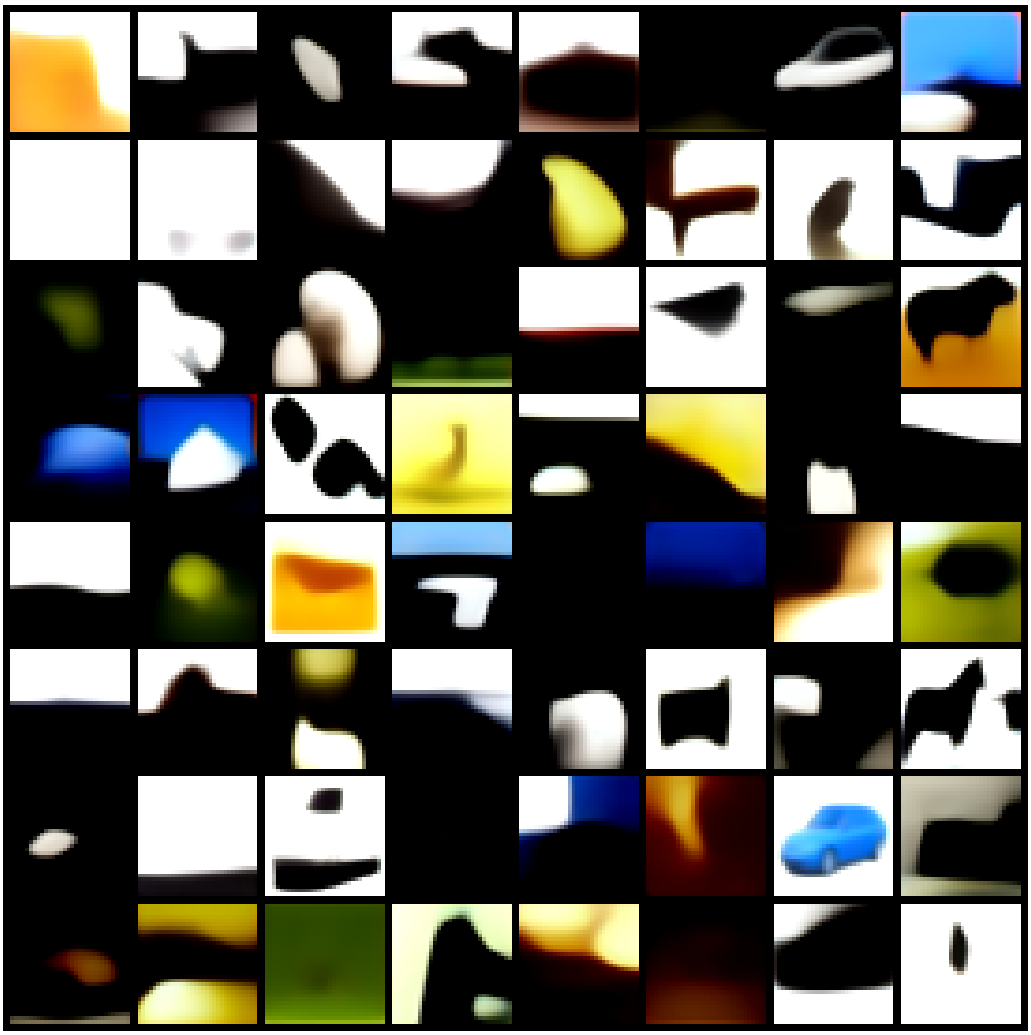

Figure 7: CIFAR-10 reconstruction: DDIM Inversion $\rightarrow$ CM

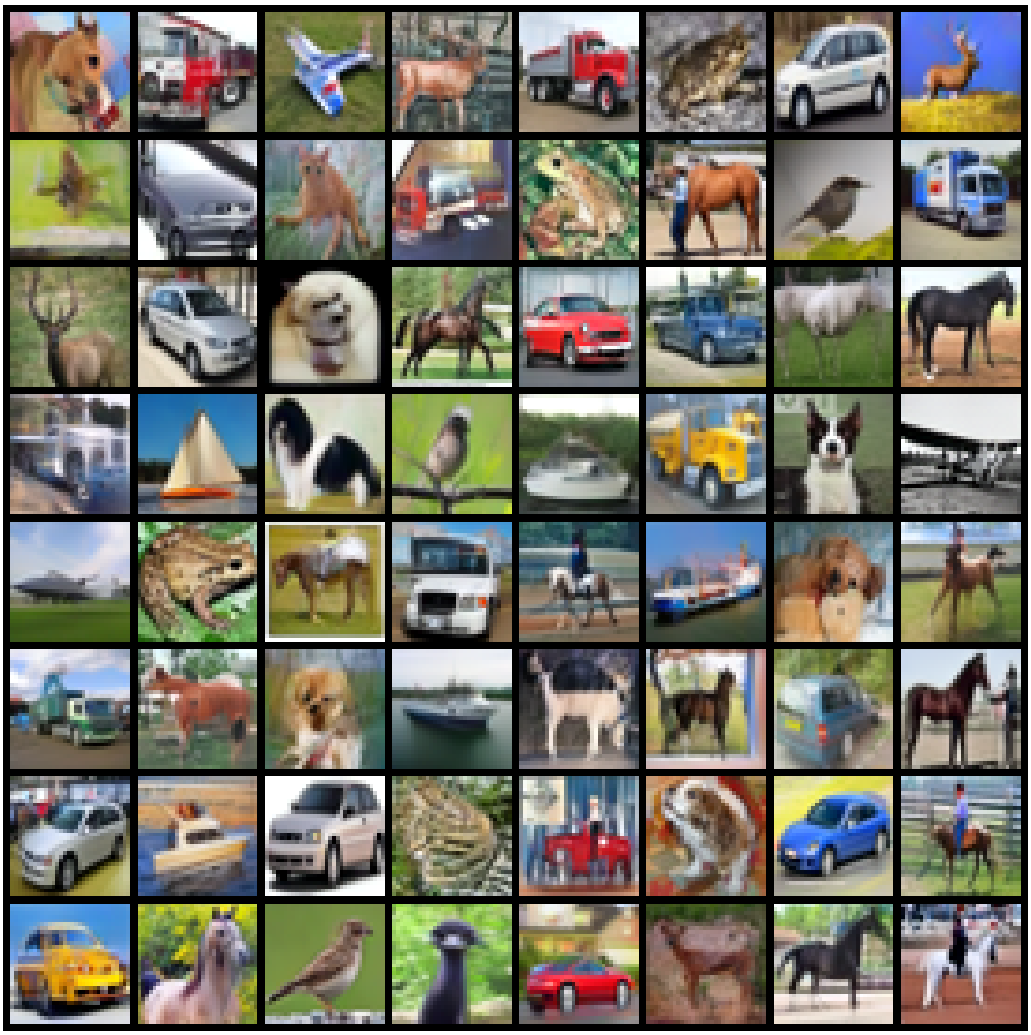

Figure 8: CIFAR-10 reconstruction: CT Inversion $\rightarrow$ CM

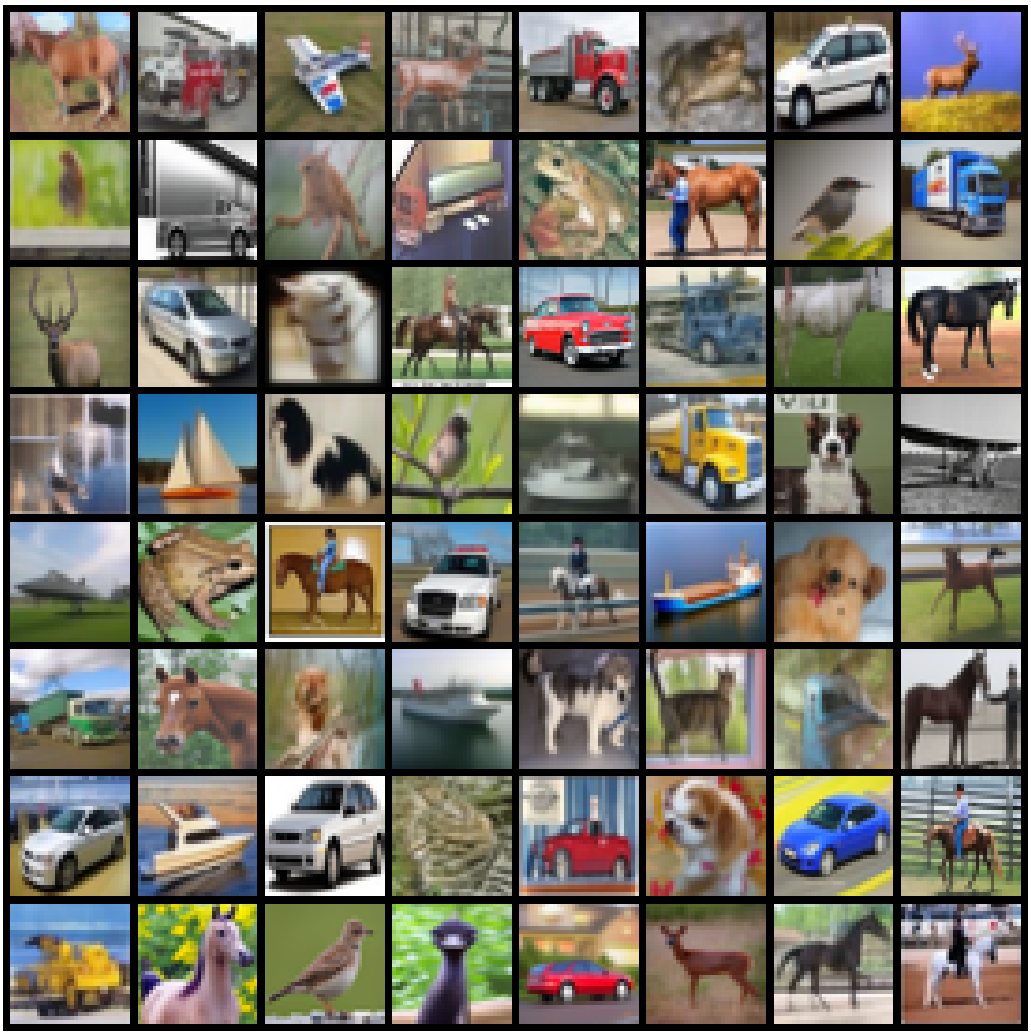

Figure 9: CIFAR-10 reconstruction: CT Inversion → DM