ACT-Diffusion: Efficient Adversarial Consistency Training for One-step Diffusion Models

Supplementary Material

A. Architecture and Experiment settings

Architecture For the consistency model architecture, we employ a structure similar to that of DDPM [19], with the exception of altering the corresponding embeddings to continuous time. We utilize the Python library diffusers [49]. In terms of the discriminator, we employ the downsampling structure in the DDPM, preserving it up to the mid-block. Subsequently, a linear layer is added to map it to $\mathbb{R}$. Additionally, the layers-per-block parameter is set to 150% of that in the consistency model, with all other parameters remaining the same. The parameters passed to the UNet2DModel are listed in Tab. A1. B=128. In the context of block type, 'D' represents DownBlock2D, 'A' stands for either AttnDownBlock2D or AttnUpBlock2D, and 'U' means UpBlock2D.

	CIFAR10	ImageNet 64×64	LSUN Cat 256×256
layers_per_block	2	2	2
block_out_channels	(1B,1B,2B,2B)	(1B,2B,2B,4B,4B)	(1B,1B,2B,2B,4B,4B)
down_block_types	DADD	DDADD	DDDDAD
up_block_types	UUAU	UUAUU	UAUUUU
attention_head_dim	8	16	16

Table A1. The parameters passed to the UNet2DModel. For those not listed, the default settings from the diffusers library are used.

Experiment settings In this section, we report the configuration of various hyperparameters within our experimental framework. Tab. A2 provides a summary of the experimental setup. Unless otherwise specified, the learning rate for both the consistency model and the discriminator is identical. The experiments conducted during the ablation study (Sec. 4.3), maintain consistency with the settings outlined in this table, with the exception of the parameters specifically varied for the ablation study. Additionally, when employing the ProjectedGAN as the discriminator, the learning rate of discriminator is set to 0.002, with w and w_{mid} values at 0.1.

Metrics The metrics used are IS, FID, Improved Precision and Improved Recall. The Inception Score (IS), introduced in [40], assesses a model's ability to generate convincing images of distinct ImageNet classes and capture the overall class distribution. However, it has a limitation in that it doesn't incentivize capturing the full distribution or the diversity within classes, leading to models with high IS even if they only memorize a small portion of the dataset, as noted in [2]. To address the need for a metric that better reflects diversity, the Fréchet Inception Distance (FID) was introduced in [18]. This metric is argued to align more closely with human judgment than IS, and it quantifies the similarity between two image distributions in the latent space of Inception-V3 as detailed in [5]. Additionally, [27] developed Improved Precision and Recall metrics that evaluate the fidelity of generated samples by determining the proportion that aligns with the data manifold (precision) and the diversity by the proportion of real samples that are represented in the generated sample manifold (recall).

Hyperparameter	CIFAR10	ImageNet 64×64	LSUN Cat 256×256
Discriminator	DDPM	DDPM	DDPM
Learning rate	1e-4	5e-5	1e-5
Batch size	80	320	320
μ_0	0.9	0.95	0.95
s_0	2	2	2
s_1	150	200	150
w_{mid}	0.3	0.2	0.1
w	0.3	0.6	0.6
I_{gp}	16	16	16
w_{gp}	10	10	10
τ	0.55	-	-
μ_p	0.93	-	-
p_r	0.05	-	-
Training iterations	300k	400k	165k
Mixed-Precision	No	Yes	Yes
Number of GPUs	1×RTX 3090	4×A100	8×A100

Table A2. Summary of the experimental setup.

B. Details of the Proof for Theorem 3.1

Details for Eq. (6):

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_{t_k},\boldsymbol{y}_{t_k}\sim\gamma^*}[\|\boldsymbol{f}(\boldsymbol{x}_{t_k},t_k,\boldsymbol{\theta})-\boldsymbol{g}(\boldsymbol{y}_{t_k},t_k)\|]\\
=&\mathbb{E}_{\boldsymbol{x}_{t_k},\boldsymbol{y}_{t_k}\sim\gamma^*}[\|\boldsymbol{g}(\boldsymbol{y}_{t_k},t_k)-\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})\\
&\qquad+\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})-\boldsymbol{f}(\boldsymbol{x}_{t_k},t_k,\boldsymbol{\theta})\|]\\
\leq&\mathbb{E}_{\boldsymbol{x}_{t_k},\boldsymbol{y}_{t_k}\sim\gamma^*}[\|\boldsymbol{g}(\boldsymbol{y}_{t_k},t_k)-\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})\|\\
&\qquad+\|\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})-\boldsymbol{f}(\boldsymbol{x}_{t_k},t_k,\boldsymbol{\theta})\|]\\
\overset{(i)}{\leq}&\mathbb{E}_{\boldsymbol{x}_{t_k},\boldsymbol{y}_{t_k}\sim\gamma^*}[\|\boldsymbol{g}(\boldsymbol{y}_{t_k},t_k)-\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})\|\\
&\qquad+L\|\boldsymbol{y}_{t_k}-\boldsymbol{x}_{t_k}\|]\\
=&\mathbb{E}_{\boldsymbol{x}_{t_k},\boldsymbol{y}_{t_k}\sim\gamma^*}[\|\boldsymbol{g}(\boldsymbol{y}_{t_k},t_k)-\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})\|]\\
&\qquad+L\mathbb{E}_{\boldsymbol{x}_{t_k},\boldsymbol{y}_{t_k}\sim\gamma^*}[\|\boldsymbol{y}_{t_k}-\boldsymbol{x}_{t_k}\|]\\
=&\mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{g}(\boldsymbol{y}_{t_k},t_k)-\boldsymbol{f}(\boldsymbol{y}_{t_k},t_k,\boldsymbol{\theta})\|]+L\mathcal{W}[q_{t_k},p_{t_k}].
\end{aligned}
$$

Here, (i) holds because $\boldsymbol{f}$ satisfies the Lipschitz condition.

Details for Eq. (7):

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{g}(\boldsymbol{y}_{t_k}, t_k) - \boldsymbol{f}(\boldsymbol{y}_{t_k}, t_k, \boldsymbol{\theta})\|] \\
\overset{(i)}{=}&\mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{g}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}) - \boldsymbol{f}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) \\
&\quad + \boldsymbol{f}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{y}^{\phi}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) \\
&\qquad + \boldsymbol{f}(\boldsymbol{y}^{\phi}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{y}_{t_k}, t_k, \boldsymbol{\theta})\|] \\
\leq&\mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{g}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}) - \boldsymbol{f}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta})\|] \\
&\qquad + \mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{f}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{y}^{\phi}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta})\|] \\
&\qquad + \mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{f}(\boldsymbol{y}^{\phi}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{y}_{t_k}, t_k, \boldsymbol{\theta})\|] \\
\overset{(ii)}{\leq}&\mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{g}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}) - \boldsymbol{f}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta})\|] \\
&\qquad + L\|\boldsymbol{y}_{t_{k-1}} - \boldsymbol{y}^{\phi}_{t_{k-1}}\| \\
&\qquad + \mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{f}(\boldsymbol{y}^{\phi}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{y}_{t_k}, t_k, \boldsymbol{\theta})\|] \\
\overset{(iii)}{=}&\mathbb{E}_{\boldsymbol{y}_{t_{k-1}}\sim p_{t_{k-1}}}[\|\boldsymbol{g}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}) - \boldsymbol{f}(\boldsymbol{y}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta})\|] \\
&\qquad + L(t_{t_k} - t_{k-1})O(t_{t_k} - t_{k-1}) \\
&\qquad + \mathbb{E}_{\boldsymbol{y}_{t_k}\sim p_{t_k}}[\|\boldsymbol{f}(\boldsymbol{y}^{\phi}_{t_{k-1}}, t_{k-1}, \boldsymbol{\theta}) - \boldsymbol{f}(\boldsymbol{y}_{t_k}, t_k, \boldsymbol{\theta})\|]
\end{aligned}
$$

Here, (i) holds because $\boldsymbol{g}$ is a consistency function, with $\boldsymbol{g}(\boldsymbol{y}_{t_k}, t_k) = \boldsymbol{g}(\boldsymbol{y}_{t_{k-1}}, t_{k-1})$. (ii) holds because $\boldsymbol{f}$ satisfies the Lipschitz condition. (iii) holds because Φ is an Euler solver, hence $\|\boldsymbol{y}_{t_{k-1}} - \boldsymbol{y}^{\phi}_{t_{k-1}}\|$ does not exceed the truncation error $O((t_n - t_{n-1})^2)$.

C. Conditional Discriminator

Theorem C.1. *Given a generator $G(\boldsymbol{z}, \boldsymbol{x}_t, t)$ and a discriminator $D(\boldsymbol{x}_0, \boldsymbol{x}_t, t)$. The distribution of optimal solution of $G(\cdot, \boldsymbol{x}_t, t)$ for the problem Eq. (11) is $p_g(\cdot|\boldsymbol{x}_t) = p(\cdot|\boldsymbol{x}_t)$, where $p_g(\cdot|\boldsymbol{x}_t)$ is the sample distribution of $G(\boldsymbol{z}, \boldsymbol{x}_t, t), z \sim p_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x}_t)$. $p_{\boldsymbol{z}}(\cdot|\boldsymbol{x}_t)$ is a normal distribution. $\boldsymbol{x}_t \sim p_t$, and $\boldsymbol{x}_0 \sim p_0$. p_t is the marginal distribution of a diffusion process.*

$$
\begin{aligned}
\min_G \max_D V(G, D) &= \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim p(\boldsymbol{x}_0, \boldsymbol{x}_t)}[\log D(\boldsymbol{x}_0, \boldsymbol{x}_t)] \\
&\quad + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x}_t), \boldsymbol{x}_t\sim p_t}[\log(1 - D(G(\boldsymbol{z}, \boldsymbol{x}_t, t), \boldsymbol{x}_t))]
\end{aligned}
\tag{11}
$$

Proof. By expressing Eq. (11) in integral form, we have the following equation:

$$
\begin{aligned}
&\iint_{\boldsymbol{x}_0, \boldsymbol{x}_t} p(\boldsymbol{x}_0, \boldsymbol{x}_t) \log(D(\boldsymbol{x}_0, \boldsymbol{x}_t)) d\boldsymbol{x}_0 d\boldsymbol{x}_t \\
&\quad + \iint_{\boldsymbol{z}, \boldsymbol{x}_t} p_{\boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{x}_t) \log(1 - D(G(\boldsymbol{z}, \boldsymbol{x}_t), \boldsymbol{x}_t)) d\boldsymbol{z} d\boldsymbol{x}_t \\
=& \int_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t) \left(\int_{\boldsymbol{x}_0} p(\boldsymbol{x}_0|\boldsymbol{x}_t) \log(D(\boldsymbol{x}_0, \boldsymbol{x}_t)) d\boldsymbol{x}_0 \right. \\
&\quad \left. + \int_{\boldsymbol{z}} p_{\boldsymbol{z}}(\boldsymbol{z}|\boldsymbol{x}_t) \log(1 - D(G(\boldsymbol{z}, \boldsymbol{x}_t), \boldsymbol{x}_t)) d\boldsymbol{z} \right) d\boldsymbol{x}_t \\
=&\mathbb{E}_{\boldsymbol{x}_t\sim p_t}\left[\int_{\boldsymbol{x}_0} p(\boldsymbol{x}_0|\boldsymbol{x}_t) \log(D(\boldsymbol{x}_0, \boldsymbol{x}_t)) \right. \\
&\quad + p_g(\boldsymbol{x}_0|\boldsymbol{x}_t) \log(1 - D(\boldsymbol{x}_0, \boldsymbol{x}_t)) d\boldsymbol{x}_0]
\end{aligned}
$$

The optimal D is:

$$
D^*_G = \frac{p(\boldsymbol{x}_0|\boldsymbol{x}_t)}{p(\boldsymbol{x}_0|\boldsymbol{x}_t) + p_g(\boldsymbol{x}_0|\boldsymbol{x}_t)}
$$

Substituting D^* into V, we obtain the following equation:

$$
\begin{aligned}
&\max_D V(G, D) \\
=&\mathbb{E}_{\boldsymbol{x}_t\sim p_t}\left[\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\log \frac{p(\boldsymbol{x}_0|\boldsymbol{x}_t)}{p(\boldsymbol{x}_0|\boldsymbol{x}_t) + p_g(\boldsymbol{x}_0|\boldsymbol{x}_t)}\right]\right. \\
&\quad \left. + \mathbb{E}_{\boldsymbol{x}_0\sim p_g(\boldsymbol{x}_0|\boldsymbol{x}_t)} \log\left[\frac{p_g(\boldsymbol{x}_0|\boldsymbol{x}_t)}{p(\boldsymbol{x}_0|\boldsymbol{x}_t) + p_g(\boldsymbol{x}_0|\boldsymbol{x}_t)}\right]\right] \\
=&\mathbb{E}_{\boldsymbol{x}_t\sim p_t}[-\log 4 + 2\mathit{JSD}(p_t(\cdot|\boldsymbol{x}_t)||p_g(\cdot|\boldsymbol{x}_t))]
\end{aligned}
$$

In the aforementioned equation, *JSD* represents the Jensen-Shannon divergence. The equation holds true only when $p_g(\cdot|\boldsymbol{x}_t) = p(\cdot|\boldsymbol{x}_t)$. This concludes the proof. □

D. ACT-Aug

In this section, we will provide the details of ACT-Aug. The differences from ACT are highlighted in red. The algorithm is listed in Algorithm 2.

E. More Experiment Results

Zero-shot Image Inpainting An important capability of consistency models is zero-shot image inpainting. This depends on the properties of the diffusion process and $\mathcal{L}_{CT}$. Given that we introduce a discriminator during the training process, does this impact the properties of consistency models? We demonstrate the results of inpainting in Fig. E3. We employ the algorithm consistent with [45]. It can be seen that ACT still retains the capabilities of consistency models.

We further display the sampling results from the conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}, \boldsymbol{x}_0 \sim p_0, \boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$ on ImageNet 64×64. k ranges from 0 to N, with 10 equidistant points. It can be observed that the sampling results of t_k and t_{k-1} exhibit significant similarity, which further substantiates that ACT does not disrupt the properties of $\mathcal{L}_{CT}$ and consistency models.

Generation Visualization on Conditional Trajectory In this section, we demonstrate samples generated from the conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}$ on ImageNet 64×64, further illustrating that our method preserves the properties of

Algorithm 2 Adversarial Consistency Training with Augmentation

```
Input: dataset $\mathcal{D}$, initial consistency model parameter $\theta_g$, discriminator $\theta_d$, step schedule $N(\cdot)$, EMA decay rate schedule $\mu(\cdot)$, optimizer opt$(\cdot,\cdot)$, discriminator with augmentation $D_{aug}(\cdot,\cdot,\cdot,\theta_d)$, adversarial rate schedule $\lambda(\cdot)$, gradient penalty weight $w_{gp}$, gradient penalty interval $I_{gp}$, gradient penalty threshold $\tau$, augmentation probability update rate $p_r$
$\boldsymbol{\theta}_g^- \leftarrow \boldsymbol{\theta}$, $k \leftarrow 0$, $p_{aug} \leftarrow 0$ and $\mathcal{L}_{gp}^- = \tau$
repeat
    Sample $\boldsymbol{x} \sim \mathcal{D}$, and $n \sim \mathcal{U}[\![1, N(k)]\!]$
    Sample $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$    ▷ Train Consistency Model
    $\mathcal{L}_{CT} \leftarrow$
        $d(\boldsymbol{f}(\boldsymbol{x} + t_{n+1}\boldsymbol{z}, t_{n+1}, \boldsymbol{\theta}_g), \boldsymbol{f}(\boldsymbol{x} + t_n\boldsymbol{z}, t_n, \boldsymbol{\theta}_g^-))$
    $\mathcal{L}_G \leftarrow \log(1-$
        $D_{aug}(\boldsymbol{f}(\boldsymbol{x} + t_{n+1}\boldsymbol{z}, t_{n+1}, p_{aug}, \boldsymbol{\theta}_g), t_{n+1}, \boldsymbol{\theta}_d))$
    $\mathcal{L}_f \leftarrow (1 - \lambda_{N(k)}(n+1))\mathcal{L}_{CT} + \lambda_{N(k)}(n+1)\mathcal{L}_G$
    $\boldsymbol{\theta}_g \leftarrow$ opt$(\boldsymbol{\theta}_g, \nabla_{\boldsymbol{\theta}_g}(\mathcal{L}_f))$
    $\boldsymbol{\theta}_g^- \leftarrow$ stopgrad$(\mu(k)\boldsymbol{\theta}_g^- + (1 - \mu(k))\boldsymbol{\theta}_g)$

    Sample $\boldsymbol{x}_g \sim \mathcal{D}$, $\boldsymbol{x}_r \sim \mathcal{D}$, and $n \sim \mathcal{U}[\![1, N(k)]\!]$
    Sample $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I})$    ▷ Train Discriminator
    $\mathcal{L}_D \leftarrow -\log(D_{aug}(\boldsymbol{x}_r, t_{n+1}, p_{aug}, \boldsymbol{\theta}_d))$
        $-\log(1 - D_{aug}(\boldsymbol{f}(\boldsymbol{x}_g + t_{n+1}\boldsymbol{z}, t_{n+1}, p_{aug}, \boldsymbol{\theta}_d))$
    $\mathcal{L}_{gp} \leftarrow w_{gp}[k \mod I_{gp} = 0]*$
        $\|\nabla_{\boldsymbol{x}_r} D_{aug}(\boldsymbol{x}_r, t_{n+1}, p_{aug}, \boldsymbol{\theta}_d)\|^2$
    $\mathcal{L}_d \leftarrow \lambda_{N(k)}(n+1)\mathcal{L}_D + \lambda_{N(k)}(n+1)\mathcal{L}_{gp}$
    $\boldsymbol{\theta}_d \leftarrow$ opt$(\boldsymbol{\theta}_d, \nabla_{\boldsymbol{\theta}_d}(\mathcal{L}_d))$
    if $k \mod I_{gp} = 0$ then
        $p_{aug} \leftarrow$
            Clip$_{[0,1]}(p_{aug} + 2([\mathcal{L}_{gp}^- >= \tau] - 0.5)p_r)$
        $\mathcal{L}_{gp}^- = \mu_p\mathcal{L}_{gp}^- + (1 - \mu_p)\mathcal{L}_{gp}$
    end if
    $k \leftarrow k + 1$
until convergence
```

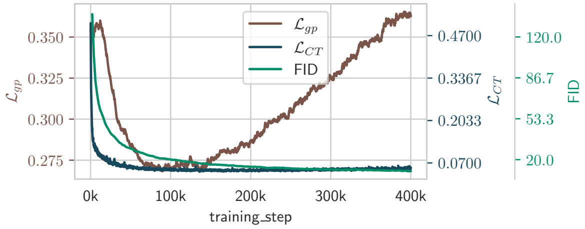


Figure E1. $\mathcal{L}_{gp}$, $\mathcal{L}_{CT}$, and FID of ACT on ImageNet 64x64 ($w_{mid=0.2}, w = 0.6$, a suitable parameter set. Under these parameters, all three metrics demonstrate stability).

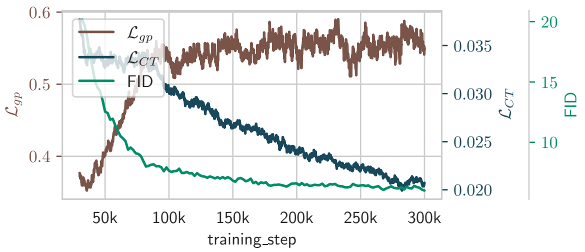


Figure E2. $\mathcal{L}_{gp}$, $\mathcal{L}_{CT}$, and FID of ACT-Aug on CIFAR10 ($\lambda_N \equiv 0.3$, a suitable parameter set. Under these parameters, all three metrics demonstrate stability).

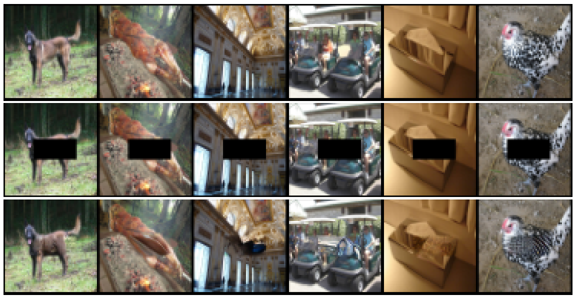


Figure E3. The results of zero-shot inpainting. **First Row:** original images; **Second Row:** masked images; **Bottom Row:** inpainted images.

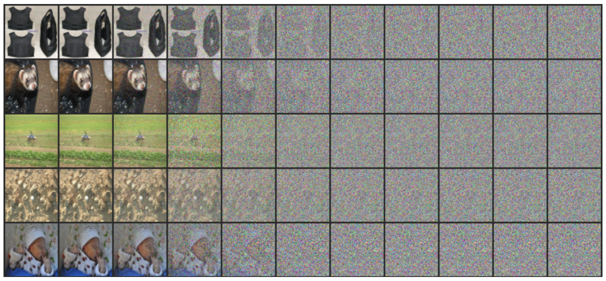


Figure E4. The conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}$ (ImageNet 64×64).

consistency training. Fig. E4 shows the conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}$, while Fig. E5 displays the samples generated from the conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}$. It can be observed that there is a high degree of similarity between adjacent t values, further validating that our method retains the properties of $\mathcal{L}_{CT}$.

Examples of proper λ_N In this section, we present the stability of $\mathcal{L}_{CT}$, $\mathcal{L}_{gp}$, and the FID score of the appropriate selection of λ_N. As depicted in Fig. E1, it is observed that all three metrics exhibit stability during training. Specifically for $\mathcal{L}_{gp}$, there is an initial decreasing trend followed by an increase; however, the variation remains within a range of 0.1 until the end of training.

Fig. E2 illustrates the stability of $\mathcal{L}_{gp}$, $\mathcal{L}_{CT}$, and the FID score for ACT-Aug under the appropriate selection of λ_N. It is observed that all three metrics exhibit stability. Furthermore, when compared with ACT on CIFAR10 as shown in

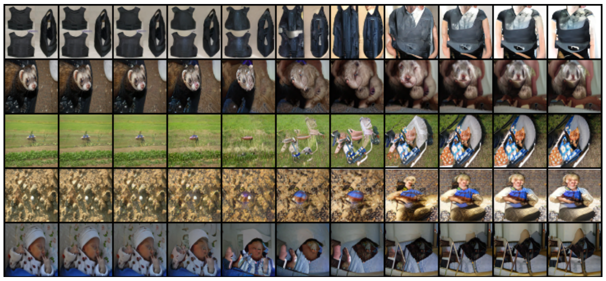
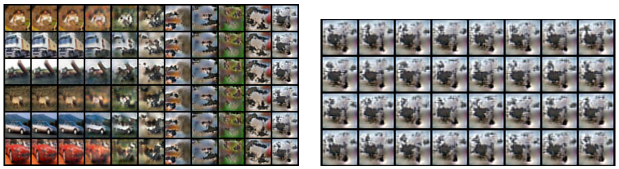


Figure E5. Generated from the conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}$ (ImageNet 64×64).



(a) Generated from the conditional trajectory $\{\boldsymbol{x}_0 + t_k\boldsymbol{z}\}$.

(b) Sampling from $T\boldsymbol{z}$.

Figure E6. Failed generations. Mode collapse when $\lambda_N \approx 1$. Experiments are conducted on the CIFAR10 dataset.

Fig. 3, $\mathcal{L}_{gp}$ is stabilized around the set $\tau = 0.55$, and both $\mathcal{L}_{CT}$ and the FID score continue to show a decreasing trend. This validates the effectiveness of the augmentation.

More samples. Fig. E6 shows failed generations on CIFAR10 dataset. Figs. E7 and E8 shows more samples on LSUN Cat 256×256 dataset.

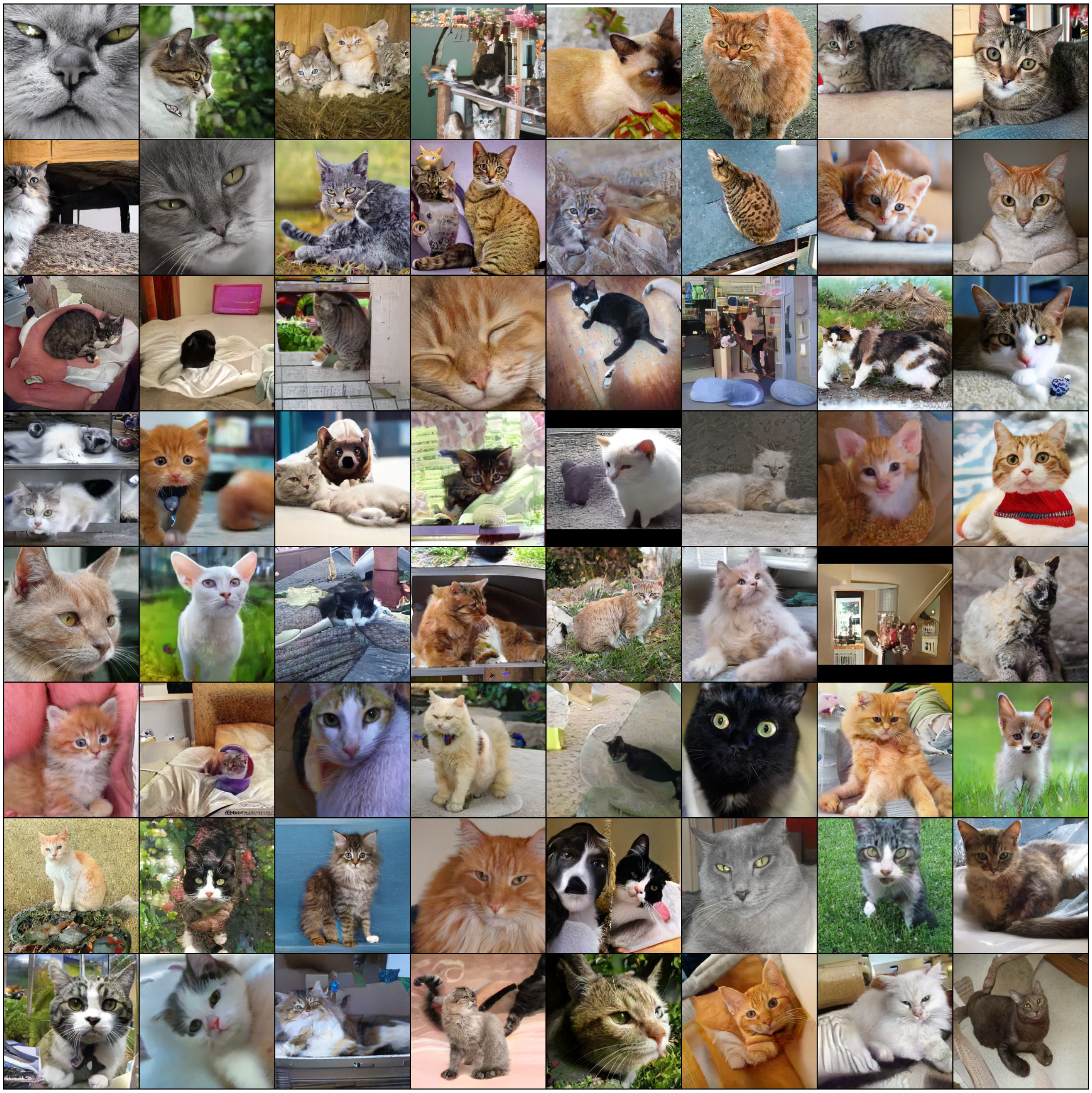


Figure E7. Generated samples (ACT Trained on LSUN Cat 256×256).

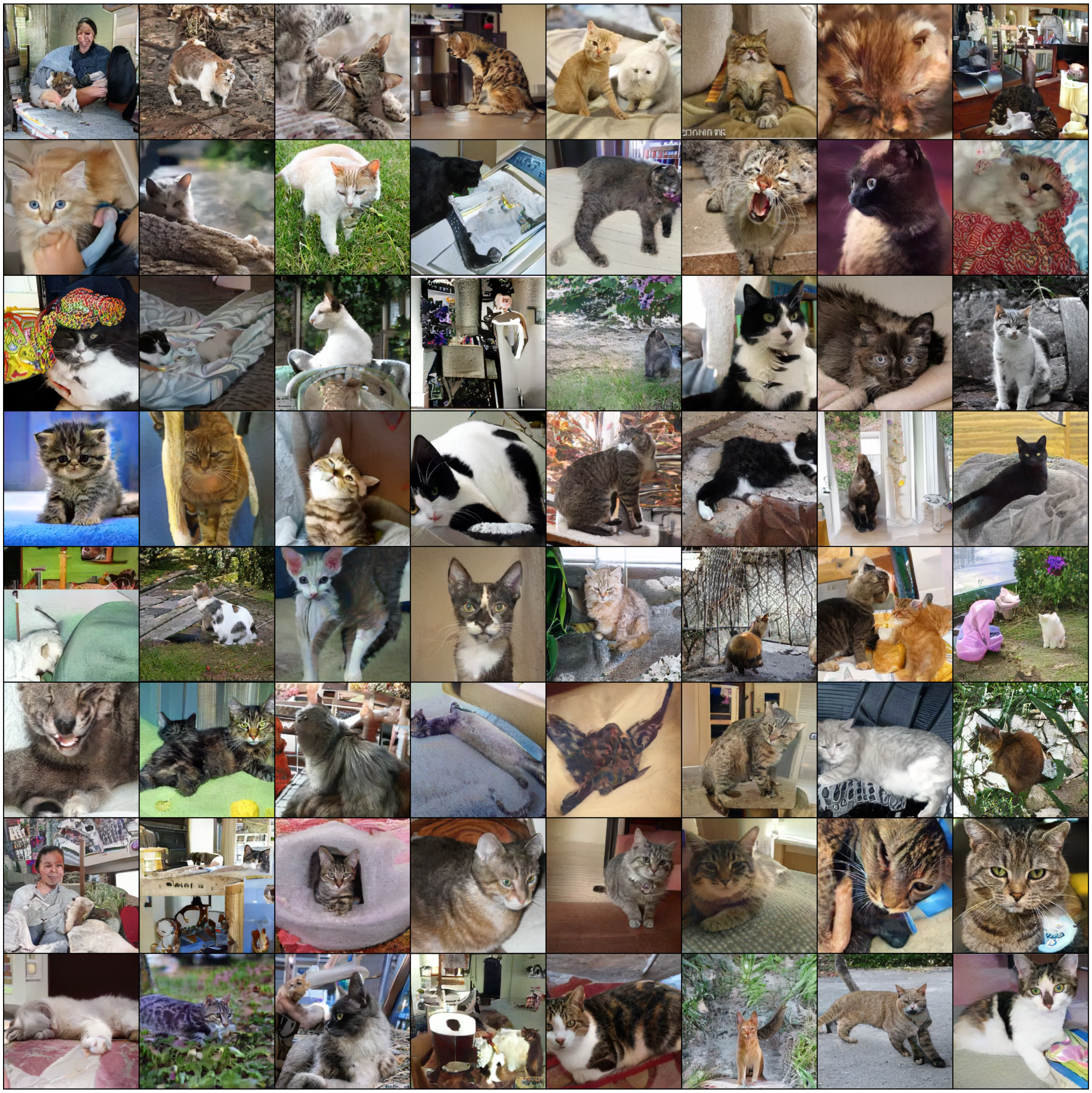


Figure E8. Generated samples (ACT Trained on LSUN Cat 256×256).