A Detailed Experiments

A.1 Network configurations

A.1.1 Generator.

Our generator follows a UNet-like architecture primarily inspired by NCSN++ (Song et al., 2021c; Xiao et al., 2022). Detailed configurations of the generator for each dataset can be found in Table 7.

A.1.2 Discriminator.

The discriminator has the same number of layers as the generator. Further details about the discriminator's structure can be found in (Xiao et al., 2022).

	CIFAR-10	STL-10	CI+MT	CE+{CH,FT,MT}
# of timesteps	4	4	4	2
# of ResNet blocks per scale	2	2	2	2
Base channels	128	128	128	96
Channel multiplier per scale	(1,2,2,2)	(1,2,2,2)	(1,2,2,2)	(1,2,2,2,4)
Attention resolutions	16	16	16	16
Latent Dimension	100	100	100	100
# of latent mapping layers	4	4	4	4
Latent embedding dimension	256	256	256	256

Table 7: Network configurations.

A.2 Training Hyperparameters

For the sake of reproducibility, we have provided a comprehensive table of tuned hyperparameters in 8. Our hyperparameters align with the baseline (Xiao et al., 2022), with minor adjustments made only to the number of epochs and the allocation of GPUs for specific datasets. In terms of training times, models for CIFAR-10 and STL-10 take 1.6 and 3.6 days, respectively, on a single GPU. For CI-MT and CE+{CH,FT,MT}, it takes 1.6 and 2 day GPU hours, correspondingly.

	CIFAR-10	STL-10	CI+MT	CE+{CH,FT,MT}
lr_G	1.6e-4	1.6e-4	1.6e-4	1.6e-4
lr_D	1.25e-4	1.25e-4	1.25e-4	1.e-4
Adam optimizer (β_1 & β_2)	0.5, 0.9	0.5, 0.9	0.5, 0.9	0.5, 0.9
EMA	0.9999	0.9999	0.999	0.999
Batch size	256	72	256	72
Lazy regularization	15	15	15	15
# of epochs	1800	800	1800	800
# of timesteps	4	4	4	2
# of GPUs	1	1	2	1
r1 gamma	0.02	0.02	0.02	0.02
Tau τ (only for RDGAN)	1e-3	1e-4	1e-3	3e-4

Table 8: Choices of hyper-parameters

A.3 Dataset Preparation

Clean Dataset: We conducted experiments on two clean datasets CIFAR-10 (32×32) and STL-10 (64×64). For training, we use 50,000 images.

Noisy Dataset:

- **CI+MI**: We resize the MNIST data to resolution of (32×32) and mix into CIFAR-10. The total samples of this dataset is 50,000 images.
- **CE+{CH,FT,MT}**: We resize the CelebHQ and CIFAR-10, FASHION MNIST, LSUN CHURCH to resolution of (64×64) and mix them together. The CelebHQ is clean and the others are outlier datasets. The noisy datasets contain 27,000 training images.

A.4 Evaluation Protocol

We measure image fidelity by Frechet inception distance (FID) (Heusel et al., 2017) and measure sample diversity by Recall metric (Kynkäänniemi et al., 2019).

FID: We compute FID between ground truth dataset and 50,000 generated images from the models

Recall: Similar to FID, we compute Recall between ground truth dataset and 50,000 generated images from the models

Outlier Ratio: We train classifier models between clean and noisy dataset and use them to classify the synthesized outliers. We firstly generate 50,0000 synthesized images and count all the outlier from them.

B Criteria for choose Ψ_i

To choose Ψ_1 and Ψ_2 for this loss function, we recommend two following criteria. First, Ψ_1 and Ψ_2 could make the trade-off between the transport map π in equation 7 and the hard constraint on the marginal distribution μ to ν in order to seek another relaxed plan that transports masses between their approximation but may sharply lower the transport cost. From the view of robustness, this relaxed plan can ignore some masses from the source distribution of which the transport cost is too high, which can be seen as outliers. Second, Ψ_1 and Ψ_2 need to be convex and differentiable so that equation 9 holds.

Two commonly used candidates for Ψ_1 and Ψ_2 are two f-divergence KL ($f(x) = x \ln x$ if $x > 0$ else $f(x) = \infty$) and χ^2 ($f(x) = (x-1)^2$ if $x > 0$ else $f(x) = \infty$). However, the convex conjugate of KL is an exponential function, making the training process for DDGAN complicated due to the dynamic of loss value between its many denoising diffusion time steps. Among the ways we tune the model, the loss functions of both generator and discriminator models keep reaching infinity.

Thus, we want a more "stable" convex conjugate function. That of χ^2 is quadratic polynomial, which does not explode when x increases like that of KL:

$$\Psi^*(x) = \begin{cases} \frac{1}{4}x^2 + x, & \text{if } x \geq -2 \\ -1, & \text{if } x < -2 \end{cases} \tag{16}$$

Remark: For any function f, its convex conjugate is always semi-continuous, and $f = f^{**}$ if and only if f is convex and lower semi-continuous (Lai & Lin, 1988). So, we can choose f^* first such that this is a non-decreasing, differentiable, and semi-continuous function. Then, we find f^{**} and check if f^{***} and f^* is equal. If f^{***} and f^*, f^{**} will be a function of which convex conjugate is f^*. Then we will check if f^{**} satisfied the first criterion to use it as Ψ_1 or Ψ_2.

C Additional Results

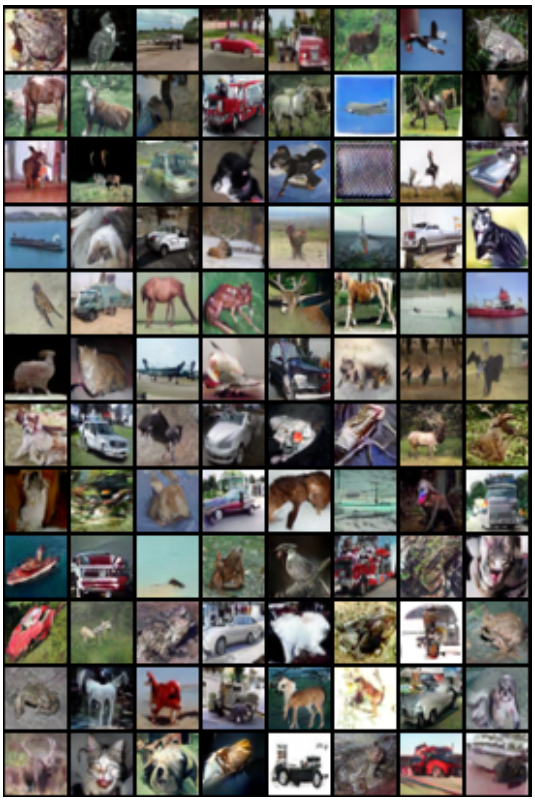
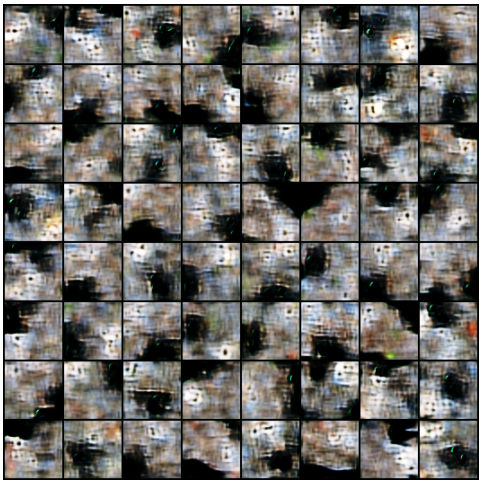
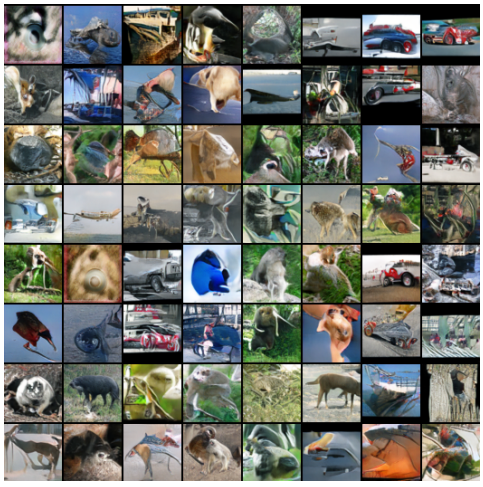


Figure 6: Non-curated CIFAR-10 qualitative images



(a) DDGAN STL-10



(b) RDGAN STL-10

Figure 7: Qualitative comparison of RDGAN and DDAN on STL-10 at epoch 300. RDGAN converges faster than DDGAN

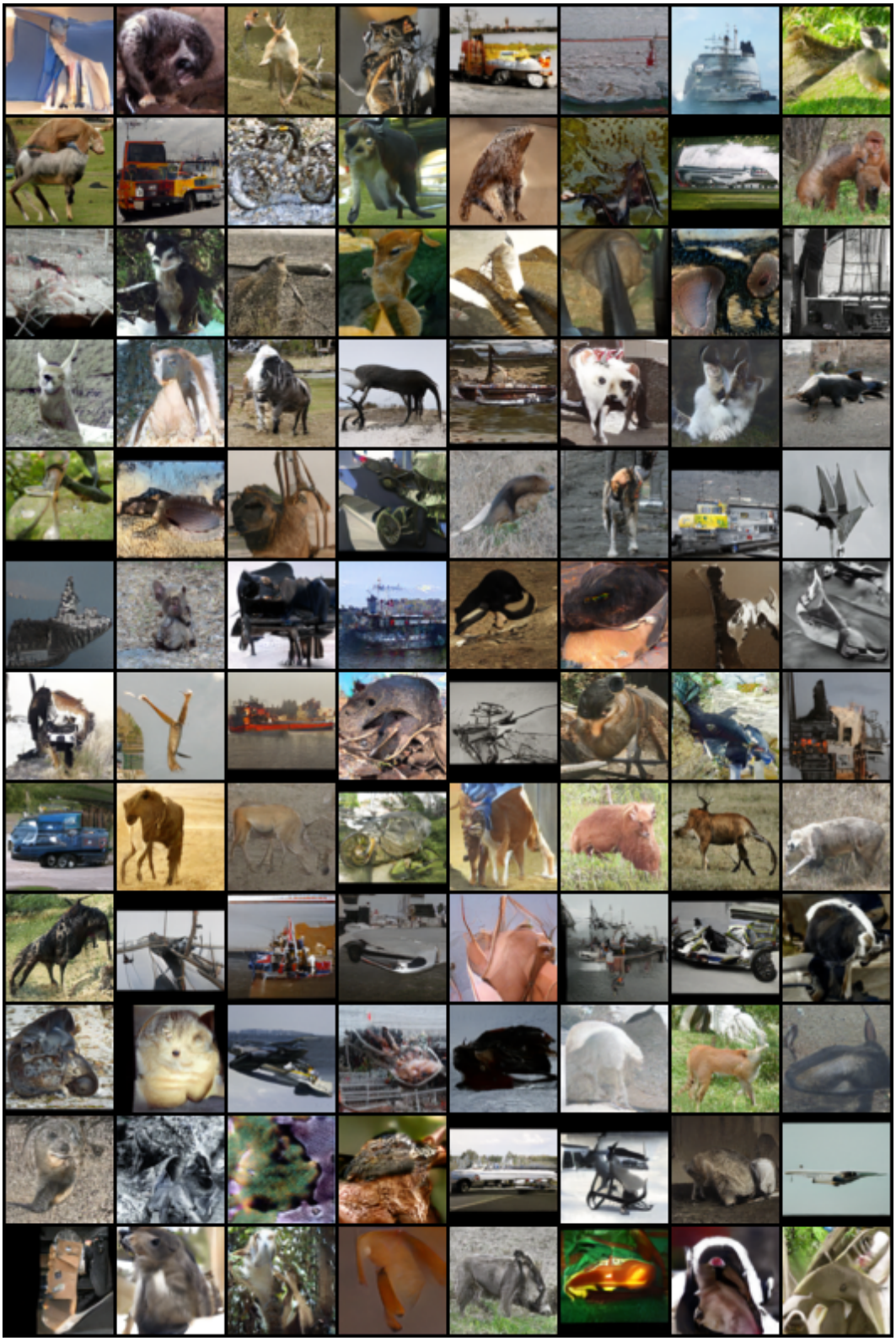


Figure 8: Non-curated STL-10 qualitative images