

610 A Resampling DLC with SEDD-absorb

611 Resampling tokens in discrete diffusion models is akin to adding noise in DDPM sampling. Such
 612 noise may have beneficial effects as it can act as an error-corrector mechanism. In [Equation 3](#),
 613 we introduced a remasking scheme of SEDD-absorb. In [Figure 8](#), we explore the effect of such a
 614 remasking scheme on the ImageNet generation FID. We find a U-shape curve where a resampling
 615 ratio that is too large cause degraded image generation quality.

616 In [Figure 9](#) we present uncurated results with $\eta \in (0., 0.01, 0.5)$. We find that η controls the image
 617 quality. η that are too small or too low produce samples that are noisy, but for different reasons.
 618 Very large η will result in a sampling with too little steps to produce good DLC. η that is too small
 619 will not allow the sampling of the DLC to correct mistakes made early in the sampling. However,
 620 contrary to classifier-free guidance, we don't find that very large η cause the model to generate weird
 621 artefacts or to reduce the diversity. This implies that remasking DLCs with SEDD-absorb or future
 622 methods could be used as a strategy to wholly replace classifier-free guidance. That said, remasking
 623 and classifier-free guidance are compatible as using one strategy do not prevent from using the other
 624 strategy.

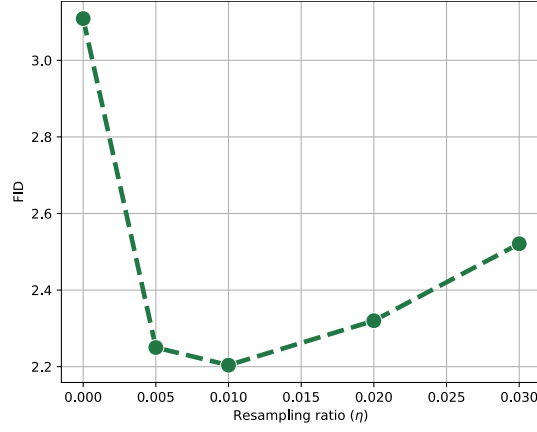


Figure 8: Effect of the resampling ratio in SEDD-absorb. Model trained on ImageNet 256×256 with DLC₅₁₂ a sequence of 512 DLC with 256 tokens each. We sample the tokens for 4096 steps and we activate the remasking for steps in $[0.3, 0.55]$. Generation without classifier-free guidance. We report the FID for several remasking ratio. We find a U-shape curve with an optimal resampling ratio of $\eta = 0.01$.

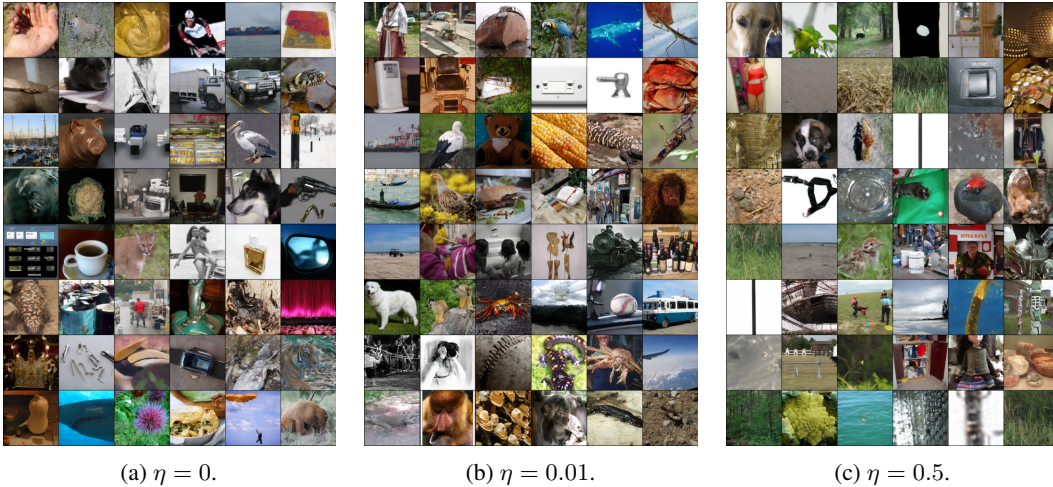


Figure 9: Qualitative resampling. Uncurated unsupervised generation for $\eta \in (0, 0.01, 0.5)$. Generation without CFG.

B Comparing DLC to Stable Diffusion

To demonstrate the benefit of DLC, we compare to the state-of-the-art open-source diffusion model, Stable Diffusion, which is both larger and trained on more data. We aim to reproduce one of our productive examples, Carbonara + Komondor, from Figure 1b. First, we generate using only a text prompt "Komondor made of Carbonara". Next, we aim to compare as closely to our average-image-embedding conditioning. We leverage IP-Adapter [Ye et al., 2023], the standard tool for image-conditioned generation. We get the CLIP embeddings from our komondor and carbonara images and generate with IP-Adapter conditioning on the average of the two image embeddings. For all generations, we use IP-adapter's recommended version Stable Diffusion 1.5, we set IP-adapter scale to 0.6, CFG to default 7.5, and generate with 100 timesteps.

We show results in Figure 10. We find that text-only conditioning is not sufficient to generate our carbonara dog, supporting our claim that text embeddings can not always sufficiently capture image semantics. Generating from the average embedding is slightly better, though lacking in diversity and failing to generate a dog at all in one case. Finally, combining both text and image conditioning allows Stable Diffusion to approach our method, though it is clearly heavily leaning towards Komondor and only changed the fur to be more pasta-like. This inability to generalize outside the komondor class may be due to the reasonably high guidance scale (7.5), but results with lower guidance scales generally failed to generate coherent images.



(a) Only text "Komondor made of Carbonara"

(b) Average of CLIP embeddings with IP-Adapter

(c) Average embedding and text

Figure 10: **Stable Diffusion can somewhat reproduce our combination of Komondor and Carbonara** but a) conditioning on purely a text prompt is insufficient and we require b) conditioning on the average CLIP embedding of an image of Komondor and Carbonara to achieve a reasonable combination of the classes, though one failure mode. Conditioning on c) both average image embedding and text prompt works best, though shows a distinct lack of diversity compared to our DLC-based method.

C Additional evidences showing inability of diffusion models to generate highly modal continuous distributions

The section 3 demonstrated that diffusion models struggle at modeling highly modal distributions. Figure 2a showcases a subset of the samples of the unconditional generation resulting for the modeling of 400 mixture. For completeness, Figure 11 showcases the full generations resulting from 16 mixtures, 100 mixtures, 400 mixtures and 800 mixtures for unconditional, oracle conditioned and GMM conditioned diffusion models. Unconditional generation observers a degrading fit as the number of modes in the dataset increase. Conditional generative model demonstrate a good fit even for very large number of modes. Interestingly, inferring the mixtures scales relatively well with the number of modes.

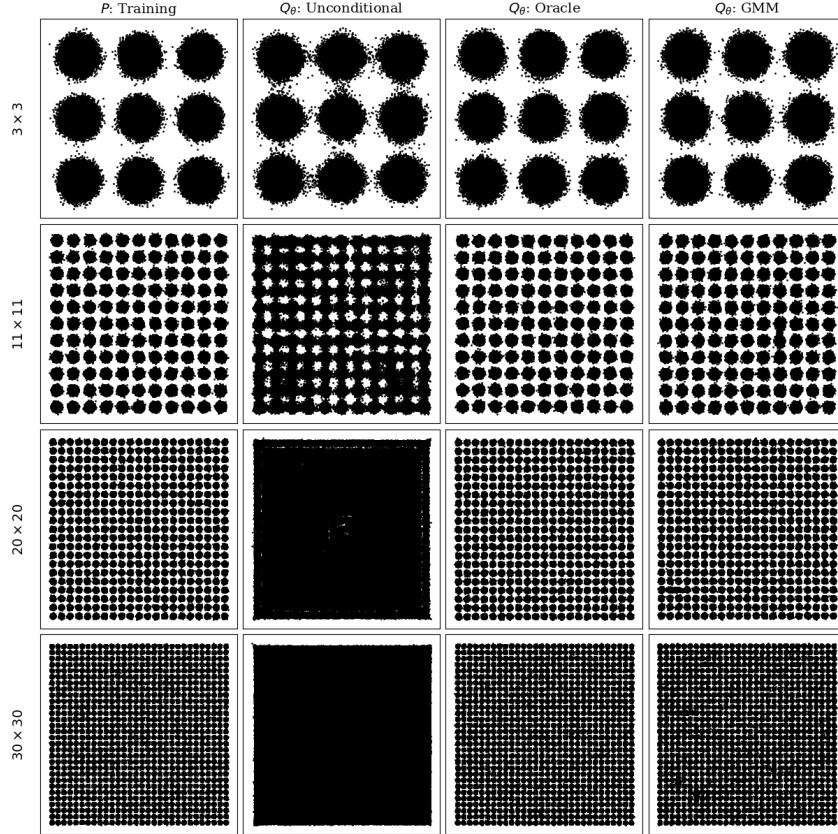


Figure 11: Comparing a $m \times m$ grid of mixture of Gaussian P with samples from model distributions Q_θ unconditional, conditioned on an *oracle* mixture index and a mixture index inferred from a Gaussian Mixture Model (GMM). Unconditional generative models cannot samples highly modals distributions. Meanwhile, conditional generative models have no problem sampling highly modal distribution. Moreover, inferring mixture indices via a GMM also scales to highly modal distributions.

653 D Evaluating Novelty of Text-to-Image samples outside ImageNet

654 Our text-and-DLC-to-image pipeline enables generation of interesting images that reflect composi-
 655 tional generalization outside the ImageNet distribution. Evaluating the novelty of generations is a
 656 fundamentally difficult task [Kingma and Gao, 2023]. To give a quantitative and qualitative sense of
 657 our pipeline’s ability, we aim to find the closest example to our generation in the ImageNet training
 658 set. To do so, we use the DINOv2 embedding space and find the K nearest neighbours in that space.
 659 We show results in Figure 12.

660 For show three types of generalization found in our model. First, we show that this model can
 661 generate samples of images that occur in ImageNet, such as a bonsai, but for which no label of
 662 bonsai exists. Such generalization showcase the utility of open-ended generation. Second, we show
 663 compositional generalization where we generate teapot in Antartical. Notably, there are no images
 664 of teapots in frozen conditions in the dataset. Our model clearly manages to learn and combine the
 665 semantics of separate foreground object and background setting. Thus, this results demonstrate that
 666 the generative image model can produce novel samples by extracting attributes in ImageNet and
 667 recomposing them in novel ways. Finally, we denote the generation of painting of flowers. While
 668 some painting of flower does exists, specific painting of the flower generated do not exists in the
 669 dataset. For example, there are not painting of the white flower in ImageNet.

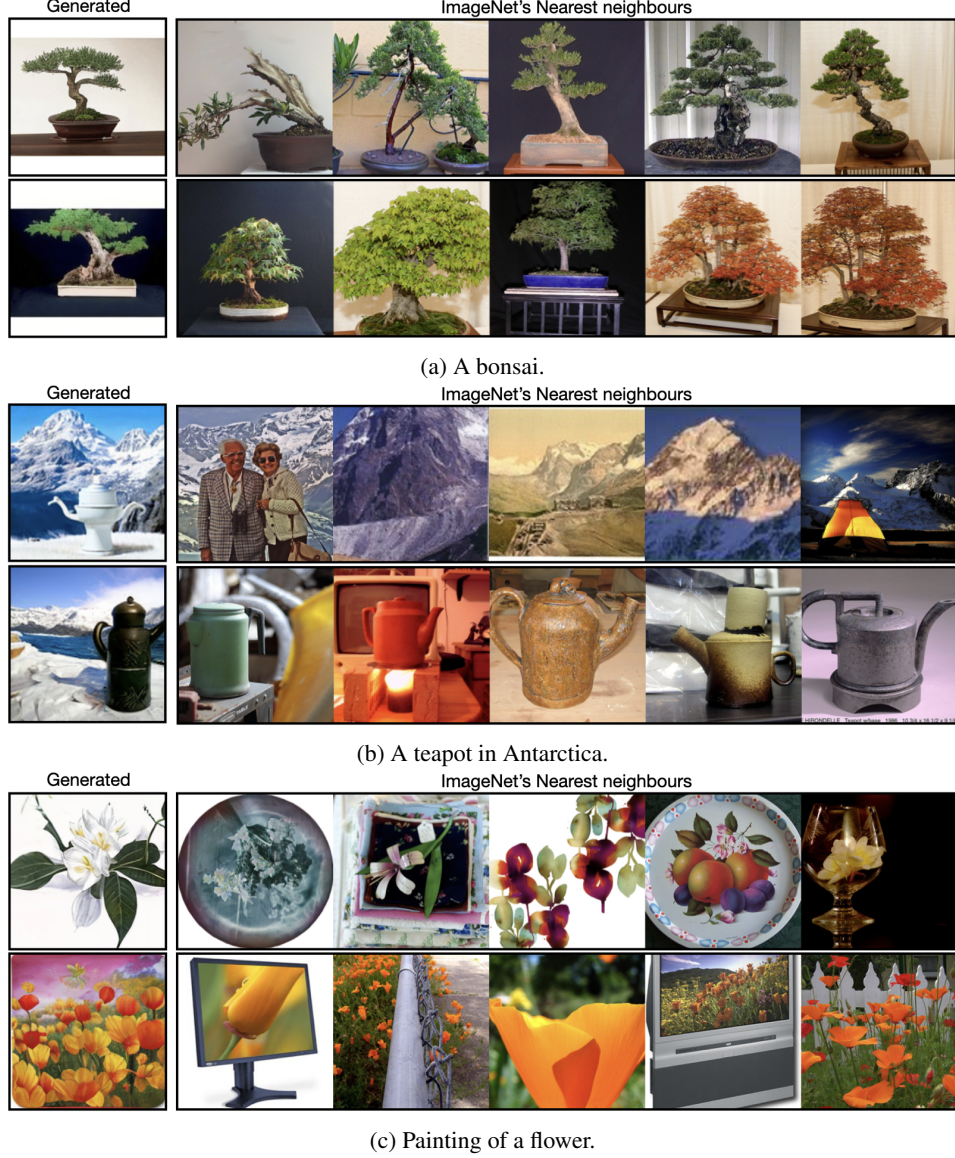


Figure 12: Text-to-image generated samples from Figure 7 and their semantic nearest neighbours in ImageNet’s training set. We find the semantic nearest neighbours with respect to the cosine similarity of DINOv2-vit/l encoding. While bonsai is not part of the labels of ImageNet, we find some *bonsai* in ImageNet’s training set that are similar to the generated samples. However, there are not *teapot in Antarctica* nor *painting of a flower* that resemble those generated by our model. This result shows the productive capability of our image generative model.

670 E Experimentation details

671 **Unconditional image generation** All models were trained with a batch size of 512 and trained
672 according to the hyper-parameters specified in Table 5. For unconditional generation of SEMs, we use
673 the Tweedie denoiser. We use remasking for generating the samples in Figure 4, Table 1 and Figure 8.
674 Otherwise, no remasking is used. We generate 50000 SEMs that are then used to generate images
675 using the image diffusion model.

676 The image diffusion models are all trained with a DiT-XL/2 model [Peebles and Xie, 2023] and
677 the hyper-parameters specified in Table 4. Following [Peebles and Xie, 2023], we use the EMA
678 VAE model from Stable-Diffusion. The generation uses DDPM [Ho et al., 2020]. We only use

N blocks	28
Hidden size	1152
Patch size	2
Num heads	16
Optimizer	AdamW
Learning rate	1e-4
Batch size	256
Weight decay	0
Num sampling steps	250
CFG scale (Table 1)	1.4
Training epochs (Table 1)	1200
Image size	256×256

Table 4: DiT-XL/2 hyper-parameters for training and sampling

N blocks	24
Hidden size	1024
Num heads	16
Optimizer	AdamW
Learning rate	3e-4
Batch size	512
Warmup	2500
Gradient clipping	1.0
Weight decay	0
Num sampling steps	4096
Resampling ratio η (Table 1)	1e-4
Training epochs (Table 1)	200

Table 5: SEDD-medium hyper-parameters for training and sampling

Model size	8B-base
# sample seen	9M
Optimizer	AdamW
Learning rate	1e-5
Total batch size	128
Grad. accum. steps	8
DLC shape	128 × 1024

Table 6: LLADA text-and-DLC hyper-parameters for fine-tuning.

679 classifier-free guidance for reporting results in Table 1. For fair comparison, we re-use the same CFG
680 scheme as DiT and apply CFG only on the first three-channel. For FID computation, we generate
681 50000 samples conditioning on the pre-sampled SEMs.

682 **Text-and-DLC fine-tuning** We fine-tune a LLADA-8B-base [Nie et al., 2025], a large diffusion
683 language model parameterized as 8B parameters Llama [Grattafiori et al., 2024] transformer [Vaswani
684 et al., 2023]. Contrary to SEDD, which predicts the concrete score, LLADA predicts the probability
685 of every tokens directly $p_\theta(x_0^i|x_t)$. The training objective to train the transformer is the cross-entropy
686 loss:

$$L(\theta) = -\mathbb{E}_{t, x_0, x_t} \left[\frac{1}{t} \sum_{i=1}^L \mathbf{1}[x_t^i = M] \log p_\theta(x_0^i|x_t) \right]. \quad (4)$$

687 For fine-tuning, we re-use the same Equation 4. However, instead of considering text tokens only
688 in our objective, we consider pairs of text and DLC tokens. The DLC tokens comes from encoded
689 images and the pairs text and images are randomly sampled from LAION [Schuhmann et al., 2022].
690 As a proof-of-concept, we randomly subsample 9M image-text pairs from LAION that are used for
691 fine-tuning.

692 For sampling the DLC, we provide a masked sequence for 128 mask tokens followed by a separator
693 token and the prompt. We follow [Nie et al., 2025] protocol with low-confidence remasking strategy.

694 F License

695 The compilation of assets used in the reproduction this work is presented in Table 7.

Asset	License	Source
ImageNet	imagenet	https://www.image-net.org/
LAION	MIT	https://github.com/LAION-AI/laion-datasets/tree/main
DinoV2	Apache 2.0	https://github.com/facebookresearch/dinov2
Fast-DiT [Jin, 2025]	CC-BY-NC-4.0	https://github.com/chuanyangjin/fast-DiT
SEDD	MIT	https://github.com/louaaron/Score-Entropy-Discrete-Diffusion
LLADA	MIT	https://github.com/ML-GSAI/LLaDA
Pytorch 2.5	Pytorch	https://github.com/pytorch/pytorch/tree/v2.5.1
Transformers	Apache 2.0	https://github.com/huggingface/transformers
This work	MIT	N/A for review.

Table 7: Compilation of assets used in the production of this work along with their license and the source location of each asset.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The main claims are that DLC:

- Improves unconditional generation of ImageNet. Demonstrated in Figure 4 and Table 1
- Enables diverse compositional generation. Demonstrated In Figure 6 and Table 3
- Enables text-to-image generation of images not in ImageNet despite training the image generator only on ImageNet (showing importance of productivity). Demonstrated in Figure 7

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The main limitation of using discrete code is the computational tradeoff as discussed in Section 5

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.