Cubic Discrete Diffusion: Discrete Visual Generation on High-Dimensional Representation Tokens

Supplementary Material

Appendix

The supplementary material includes the following additional information:

- Sec. A provides more implementation details for generation and understanding experiments.
- Sec. B presents additional experiments of CubiD on low-dimensional tokens.
- Sec. C discusses limitations.
- Sec. D showcases additional image generation results.

A. Implementation Details

A.1. Generation Experiments

Additional Training Details. We train all CubiD models on the ImageNet-1K [8] training set, consisting of 1,281,167 images across 1,000 object classes. Beyond the details provided in the main paper, we use a batch size of 2048 distributed across all GPUs. We employ mixed precision training with fp16 to reduce memory consumption and accelerate training. An exponential moving average (EMA) of model weights is maintained with momentum 0.9999 for stable evaluation. The learning rate warmup is applied for the first 100 epochs. We adopt the noise-augmented decoder from [53], which injects Gaussian noise into clean latents during decoder training to improve robustness to imperfect generative outputs.

A.2. Understanding Experiments

Training. To validate the understanding performance of our discretized tokens, we adopt the classic LLaVA [25] visual instruction tuning framework, and perform experiments with the original representations and discretized tokens. Following its standard protocol, we first perform pretrain on 558K LAION-CC-SBU [22] subset for 1 epoch, then conduct visual instruction tuning on the LLaVA-Instruct-665K dataset for 1 epoch. We use Vicuna-13B-v1.5 [55] as the language backbone and maintain all original hyperparameters, with the only modification being the replacement of continuous vision features with their quantized counterparts.

Evaluation. We evaluate on four standard benchmarks from the LLaVA evaluation suite: GQA [15] for compositional visual reasoning, TextVQA [35] for text recognition and understanding in images, POPE [23] for assessing hallucination tendencies, and MME [11] for comprehensive multimodal perception capabilities. These benchmarks collectively measure whether the quantized representations maintain the diverse understanding abilities required for multimodal tasks.

Table 6. **CubiD on low-dimensional tokens on ImageNet 512×512.** Results using DC-AE-f32c32 tokenizer producing 32-dimensional tokens.

Method	Params (B)	gFID↓	IS↑
SiT-XL [27]	0.67	2.41	131.37
USiT-H [4]	0.50	1.89	174.58
USiT-2B [4]	1.58	1.72	187.68
CubiD	0.95	**1.58**	**188.70**

Table 7. **CubiD with compressed representation tokens on ImageNet 256×256.** Features compressed from 768d to 32d.

Method	Token Dim	gFID↓	IS↑
CubiD	32	**1.55**	**296.5**

B. CubiD on Low-Dimensional Tokens

To validate the generality of CubiD beyond high-dimensional representations, we conduct experiments on traditional low-dimensional tokens.

B.1. Traditional Reconstruction-based Tokens

We employ DC-AE-f32c32 [4], a state-of-the-art autoencoder with patch size 32 that produces 32-dimensional tokens. For 512×512 images, this results in 16×16×32 discrete tokens after dimension-wise quantization, which are significantly more compact than the 32×32×768 tokens in our main experiments. As shown in Table 6, CubiD achieves 1.58 gFID and 188.7 IS on ImageNet 512×512, outperforming previous state-of-the-art methods using the same tokenizer, including USiT-2B (1.72 gFID) despite using fewer parameters. This demonstrates that our cubic discrete diffusion formulation is effective across different token dimensionalities.

B.2. Compressed Representation Tokens

To explore the generation-understanding trade-off, we investigate CubiD's performance on compressed representation tokens. We reduce the original high-dimensional features to 32 dimensions using a learned projection layer optimized for reconstruction quality. Table 7 shows that compressed 32-dimensional tokens achieve strong generation performance (1.55 gFID, 296.5 IS). While lower-dimensional spaces naturally facilitate easier generation,

this compression inevitably degrades the representation quality needed for understanding tasks. Therefore, we choose to model the original high-dimensional tokens to preserve both generation and understanding capabilities.

C. Limitations

While CubiD demonstrates the feasibility of discrete generation on high-dimensional representation tokens, several limitations remain.

Dependence on Representation Encoder. Since CubiD operates on features from a frozen pretrained encoder, the reconstruction quality sets an upper bound on generation quality. In our experiments, the reconstruction PSNR is approximately 18 dB, which limits the fine-grained details in generated images. Improving the reconstruction capability of representation autoencoders remains a valuable direction for future work.

Gap with Continuous Generation. Although discrete generation offers advantages for unified multimodal modeling through a shared cross-entropy objective, there still exists a gap compared to continuous diffusion methods such as RAE [53]. We believe this gap can be further narrowed with advances in discrete generative modeling.

Inference Efficiency. CubiD requires more generation steps than continuous diffusion models. Achieving high-quality generation typically requires hundreds to a thousand steps. Accelerating discrete diffusion inference, potentially through techniques developed for discrete language models, remains an important direction for future work.

D. More Visualization Results

Figure 6. **Uncurated samples on ImageNet 256×256 using CubiD-XXL conditioned on the specified classes.**

Figure 7. **Uncurated samples on ImageNet 256×256 using CubiD-XXL conditioned on the specified classes.**