A. Effectiveness of TexTok on Discrete Tokens

In the main paper, we demonstrate that TexTok works well with continuous tokens. In this section, we further validate the effectiveness of TexTok with Vector-Quantized (VQ) discrete tokens. We use a codebook size of 4096. As shown in Table 6, TexTok consistently delivers significant improvements over Baseline (w/o text) in reconstruction performance. As the number of tokens decreases, the performance gains are more pronounced. These results verify the effectiveness of TexTok on discrete tokens, highlighting its versatility as a universal tokenization framework. This inherent compatibility with both continuous and discrete tokens allows TexTok to seamlessly integrate with a wide range of generative models, including diffusion models, autoregressive models, and others.

tokenizer	# tokens	rFID ↓	rIS↑	PSNR↑	SSIM ↑	LPIPS↓
Baseline-32	32	7.71	84.0	15.52	0.3822	0.4524
TexTok-32		**4.11**	**141.4**	**16.52**	**0.4040**	**0.3855**
Baseline-64	64	4.34	110.1	17.11	0.4200	0.3470
TexTok-64		**2.50**	**161.8**	**18.06**	**0.4462**	**0.2933**
Baseline-128	128	2.34	139.8	18.91	0.4737	0.2476
TexTok-128		**1.76**	**167.9**	**19.96**	**0.4926**	**0.2166**
Baseline-256	256	1.45	159.4	20.67	0.5371	0.1848
TexTok-256		**1.17**	**180.3**	**21.56**	**0.5526**	**0.1594**

Table 6. **Reconstruction performance comparison of TexTok with Baseline (w/o text) using *discrete* tokens** on ImageNet 256×256. TexTok works well with discrete tokens. It consistently outperforms Baseline (w/o text) by a large margin in image reconstruction quality, achieving more pronounced gains as the number of tokens decreases.

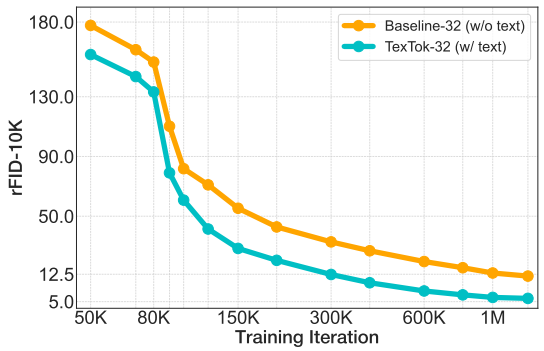

Figure 7. **Training reconstruction FID comparison of TexTok-32 *v.s.* Baseline-32 (w/o text)** on ImageNet 512×512. TexTok training is more efficient and effective, achieving faster convergence and better reconstruction quality.

B. Additional Training Analysis

In Figure 7, we further provide the training reconstruction FID comparison (evaluated on 10K samples) of TexTok-32 *v.s.* Baseline-32 (w/o text) on ImageNet 512×512. From the figure, it is clear that TexTok training is more efficient and effective, achieving faster convergence and better reconstruction quality.

C. Additional Implementation Details

We provide detailed default training hyperparameters for TexTok-N as listed below:

- ViT encoder/decoder hidden size: 768.
- ViT encoder/decoder number of layers: 12.
- ViT encoder/decoder number of heads: 12.
- ViT encoder/decoder MLP dimensions: 3072.
- ViT patch size: 8 for 256×256 image resolution and 16 for 512×512.
- Discriminator base channels: 128.
- Discriminator channel multipliers: $1, 2, 4, 4, 4, 4$ for 256×256 image resolution, and $0.5, 1, 2, 4, 4, 4, 4$ for 512×512.
- Discriminator starting iterations: $80,000$.
- Latent shape: $N \times 8$.
- Reconstruction loss weight: 1.0.
- Generator loss type: Non-saturating.
- Generator adversarial loss weight: 0.1.
- Discriminator gradient penalty: r1 with cost 10.
- Perceptual loss weight: 0.1.
- LeCAM weight: 0.0001.
- Peak learning rate: 10^{-4}.
- Learning rate schedule: linear warm up and cosine decay.
- Optimizer: Adam with $\beta_1 = 0$ and $\beta_2 = 0.99$.
- EMA model decay rate: 0.999.
- Training epochs: 270.
- Batch size: 256.

We provide detailed default training and evaluation hyperparameters for DiT as listed below:

- Patch size: 1.
- Peak learning rate: 5×10^{-4}.
- Learning rate schedule: linear warm up and cosine decay.
- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and 0.01 weight decay.
- Diffusion steps: 1000.
- Noise schedule: Linear.
- Diffusion $\beta_0 = 0.0001$, $\beta_{1000} = 0.02$.
- Training objective: v-prediction.
- Sampler: DDIM.
- Sampling steps: 250.
- Training epochs: 350.
- Batch size: 1024.

For classifier-free guidance, we adopt the guidance schedule from [12] following prior arts [13, 49]. For the model configurations and other implementation details, please refer to the original DiT paper [32].

25 Words: Black and white husky, tongue out, sits in snowy mountains, chained to a heavy anchor, thick fur coat.

75 Words: A medium-sized husky with a thick, predominantly white coat speckled with black patches sits in a snowy landscape against a backdrop of distant, snow-capped mountains. Its tongue lolls out slightly, suggesting a relaxed state, and it is tethered to a heavy anchor with a thick metal chain. The husky's fur appears soft and dense, its black markings accentuating its white body, and its expression is calm and gentle.

25 Words: Sunlit, orange-brown acoustic guitar rests against a wall near a woven fabric, showcasing smooth wood and strings.

75 Words: A sun-kissed, medium-sized acoustic guitar with a warm, sunburst finish rests against a tan stucco wall and a rustic woven fence. Its smooth, glossy body contrasts with the rough texture of the wall and the earthy tones of the fence. The guitar's dark brown fretboard and strings stand out against the vibrant orange-brown body, and a simple, elegant soundhole is centrally positioned, capturing the light. A soft fabric partially obscures the ground near the guitar.

25 Words: Fluffy silver-gray Persian cat lounges relaxed on a tan fabric couch, showcasing a plush, long-haired coat and a fluffy tail.

75 Words: A fluffy, silver-and-white Persian cat lounges comfortably on a beige, textured sofa, its long, luxurious fur creating a soft cloud-like texture. The cat's round body and plush tail are relaxed, its paws tucked gently beneath it, and its green eyes are partially visible in a somewhat pensive expression. It appears content and at ease in its domestic environment.

25 Words: A close-up of a metallic mesh microphone head, gleaming against a pink backdrop. Its textured, silver grid contrasts with the smooth, dark mic body.

75 Words: A silver dynamic microphone dominates the frame against a soft pink backdrop. Its metallic mesh grille, a tightly woven texture reflecting light, curves over the spherical head. A dark, matte handle angles out of view, suggesting its handheld size. The close-up perspective emphasizes the microphone's readiness to capture sound.

25 Words: Snow-capped mountains frame a tranquil valley, reflecting in a still, dark-blue lake under a vibrant blue sky.

75 Words: A serene valley nestled between snow-capped mountains reflects perfectly in a tranquil, deep blue lake under a vibrant azure sky. Sparse, verdant vegetation lines the valley floor, contrasting with the rugged, dark brown and gray mountain textures. Small clusters of houses dot the landscape near the water's edge, creating a picturesque scene of quiet, natural beauty.

25 Words: Brick red church with tall spire, nestled amidst green trees and a landscaped park at sunset.

75 Words: A vibrant red brick church with a tall, pointed spire stands amidst lush green trees and manicured lawns under a soft, dusky sky. Its dark gray roof contrasts with the warm brick, and arched windows line the facade, hinting at a traditional architectural style. The church is nestled in a tranquil urban setting, with a hint of city skyline in the distance.

Figure 8. **ImageNet captioning examples**. Both 25-word and 75-word captions are displayed to the right of the corresponding images.

D. ImageNet Captioning Details

We provide the prompt fed into Gemini that we used to generate the captions for images in ImageNet training and validation sets:

```
Describe an image of a {class_name} from the
ImageNet dataset in a single continuous detailed
paragraph, using no more than {word_count} words.
Include descriptions of the appearance, colors,
textures, size, environment, and any notable
features or actions. Make sure to provide a vivid
and engaging description that captures the essence
of the {class_name}. Output only the description
without any additional words or commentary and the
description must not exceed {word_count} words.
<image_bytes>
```

Figure 8 shows several captioning examples, including both 25-word and 75-word captions.

E. Additional Qualitative Results

We present additional qualitative class-conditional image generation results on ImageNet 256×256 in Figure 9. We observe that TexTok generates high-quality, semantically meaningful images with intricate fine-grained details.

We also present additional text-to-image generation results on ImageNet 256×256 in Figure 10. Note that TexTok generates photo-realistic images that accurately align with the given prompts and even share many visual details with the reference images that the model has never seen, demonstrating both high fidelity and the capability to follow the text prompts.

F. Additional Discussion on Related Work

There have been some recent efforts in image tokenization methods [29, 30, 48, 53] that aim to align image tokens with textual semantics. Approaches like LQAE [30], SPAE [48], and V2L [53] map images into tokens derived from the codebooks of large language models (LLMs), while LG-VQ [29] focuses on aligning the decoder features of image tokens with text representations. These methods are designed primarily for bridging visual and textual modalities to improve multi-modal understanding tasks. However, by aligning image tokens directly to textual semantic spaces, they often suffer from limited image reconstruction quality due to the inherent divergence between vision and language representations. Consequently, these approaches fail to achieve reasonable performance, or even do not report results, on standard image generation benchmarks, such as ImageNet class-conditional generation.

In contrast, our work introduces a novel image tokenization framework specifically designed for generative tasks. We are the first work that proposes conditioning the tokenization process directly on the text captions of images, a strategy typically reserved for the generation phase. Rather than enforcing a strict alignment between image tokens and

text captions, our method leverages descriptive captions to provide a compact semantic representation. This simplifies semantic learning, allowing more learning capacity and token space to be allocated to capture fine-grained visual details. This complementary approach significantly improves the reconstruction quality and compression rate of tokenization. Furthermore, we demonstrate that our approach achieves state-of-the-art performance on standard ImageNet conditional generation benchmarks on both 256×256 and 512×512 image resolutions, establishing our method as a distinct advancement over existing works.

Figure 9. **Qualitative class-conditional image generation results** on ImageNet 256×256. TexTok generates high-quality, semantically meaningful images with intricate fine-grained details. Results are generated with our class-conditional TexTok-256 + DiT-XL model.

Figure 10. **Qualitative text-to-image generation results** on ImageNet 256×256. For each sample, we present the reference image from the ImageNet validation set alongside the corresponding generated image, which is produced conditioned on the caption (displayed below) of the reference image. Results are generated with our text-to-image TexTok-128 + DiT-XL-T2I model. TexTok generates visually realistic images that accurately align with the given prompts, demonstrating both high fidelity and semantic relevance.