Image Tokenizer

김초은

Image Tokenizer?

Text

"an orange cat"

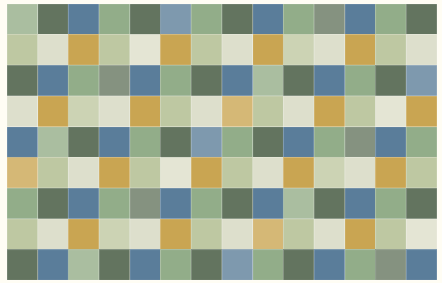
↓

[42, 318, 711]

↓

Language Model

Image



Pixels

↓

?

Continuous Vectors vs. Discrete Tokens

Continuous Vectors

$$z \in \mathbb{R}^{h \times w \times d}$$

- 실수 latent를 사용
- latent diffusion 계열에서 자주 사용
- 부드러운 복원과 고해상도 생성에 유리
- 텍스트 token처럼 categorical prediction하기는 어려움

Discrete Tokens

$$t \in \{1, \ldots, K\}^{h \times w}$$

- 정수 token ID를 사용
- AR/Masked LM을 활용하여 이미지 생성 비용을 낮춤
- 텍스트·이미지·비디오 token을 하나의 sequence로 결합 가능
- quantization error와 codebook collapse가 문제

어디에 쓰이는가?

visual token은 생성, 압축, 이해, 멀티모달 모델의 공통 interface로 쓰입니다.

AR image generation

$p(t_1, \cdots, t_n) = \prod p(t_i | t_{<i})$

Masked generation

[12, [MASK], 91, [MASK]]

Multimodal LM

Text + Image (Video) + Audio

Compression

tokens 저장/전송 → decoder 복원

Visual understanding

token 단위 interpretation

흐름

2D Spatial Patch → Lookup-Free Quantization → Compact 1D →Semantic & Concept Realignment

VQ 기반 이미지 tokenization

encoder 출력 벡터를 codebook에서 가장 가까운 벡터로 치환합니다.

$$z = [1.3, 0.1, -0.2]$$

↓

Codebook C

$$\begin{bmatrix} [\,0.2 & -0.7 & 1.1\,] \\ [\,1.5 & 0.3 & -0.4\,] \\ [\,-0.8 & 1.2 & 0.5\,] \\ [\,0.1 & 0.9 & -1.3\,] \end{bmatrix}$$

↓

nearest: c_2 → Token ID 2

장점

- 이미지를 정수 sequence로 변환
- Transformer가 language token처럼 예측
- 높은 compression ratio

문제

- 사용되지 않는 code 발생
- code를 만들기 어려움

Source: VQ-VAE / “Neural Discrete Representation Learning”, NIPS 2017

더 단순하고 확장 가능한 Quantization

거대한 lookup table 대신, 작은 factor들의 조합으로 큰 vocabulary를 만듭니다.

기존 VQ

$$C \in \mathbb{R}^{K \times d}$$

Token ID마다 학습된 code vector가 필요

K=262K, d=256 → 67M scalars

큰 vocabulary에서 비용과 collapse 문제가 커짐

Lookup-Free Quantization

$$z \in \mathbb{R}^{18} \rightarrow sign(z) \in \{-1, 1\}^{18}$$

2^{18}= 262,144 tokens

예시

[1.7, −0.3, 2.1, −1.2] → [+1, −1, +1, −1] → 1010 → ID 10

큰 codebook을 저장하지 않고 binary factor 조합으로 만든다.

Source: MAGVIT-v2 / “Language Model Beats Diffusion”, ICLR 2024

더 단순하고 확장 가능한 Quantization

거대한 lookup table 대신, 작은 factor들의 조합으로 큰 vocabulary를 만듭니다.

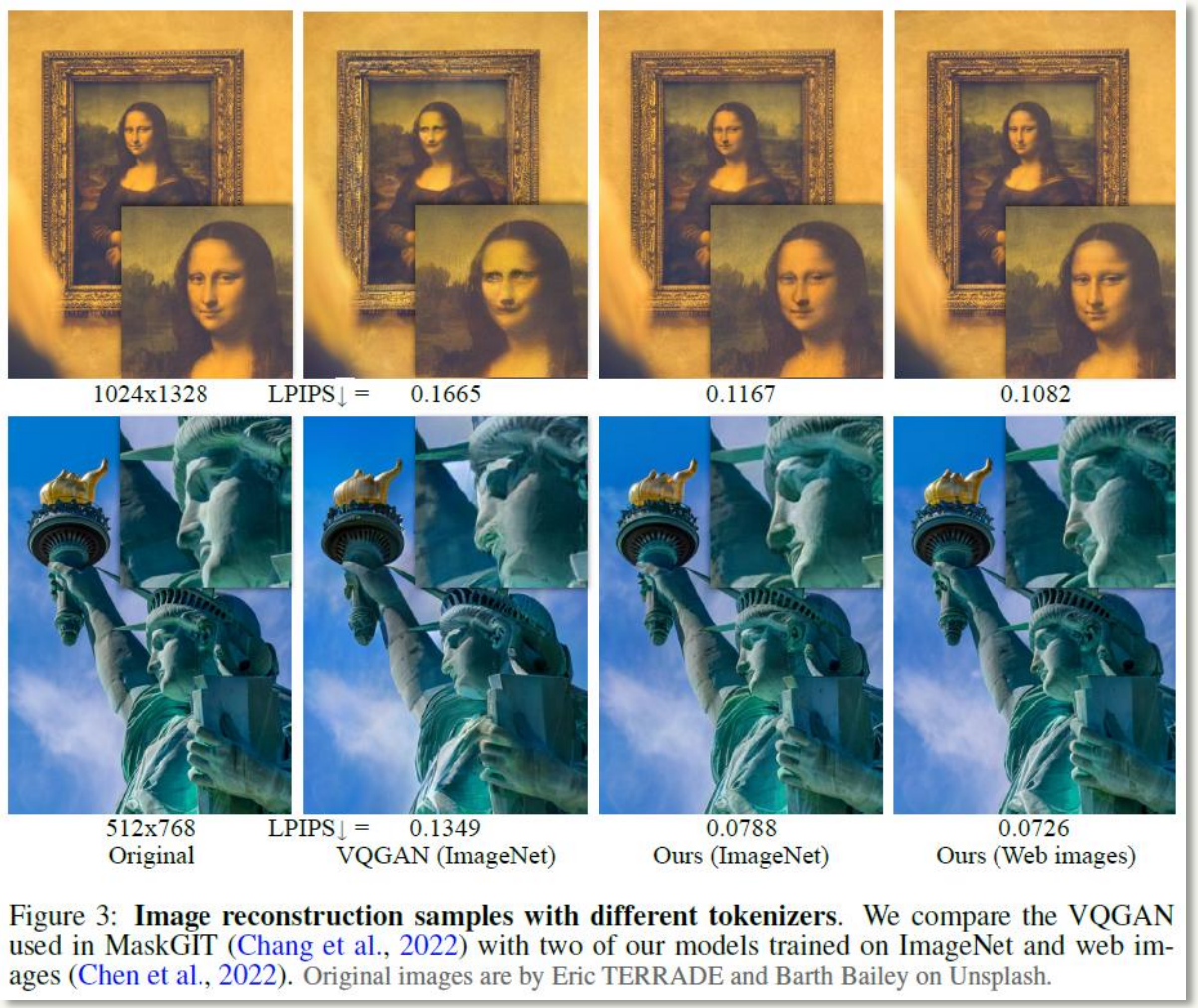


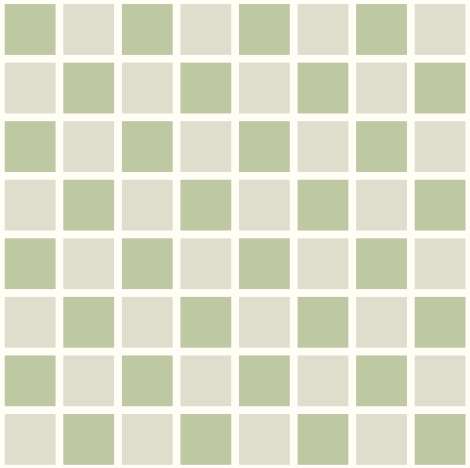
Figure 3: **Image reconstruction samples with different tokenizers**. We compare the VQGAN used in MaskGIT (Chang et al., 2022) with two of our models trained on ImageNet and web images (Chen et al., 2022). Original images are by Eric TERRADE and Barth Bailey on Unsplash.

Source: MAGVIT-v2 / "Language Model Beats Diffusion", ICLR 2024

Token 개수 줄이기

Token 개수를 줄이기 위해 2D 격자 모양에서 벗어나 이미지 전체의 중요한 정보를 골고루 흡수하도록 학습합니다.

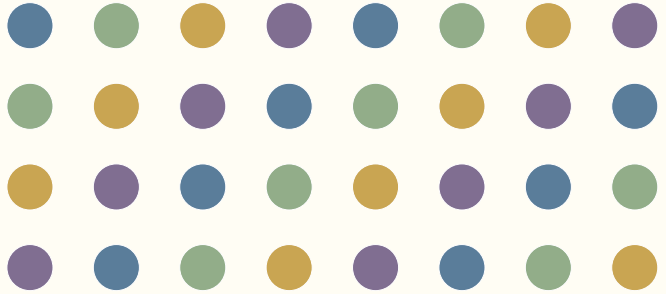
2D grid tokenizer



16 × 16 = 256 tokens

spatial grid에 고정 대응

1D compact tokenizer



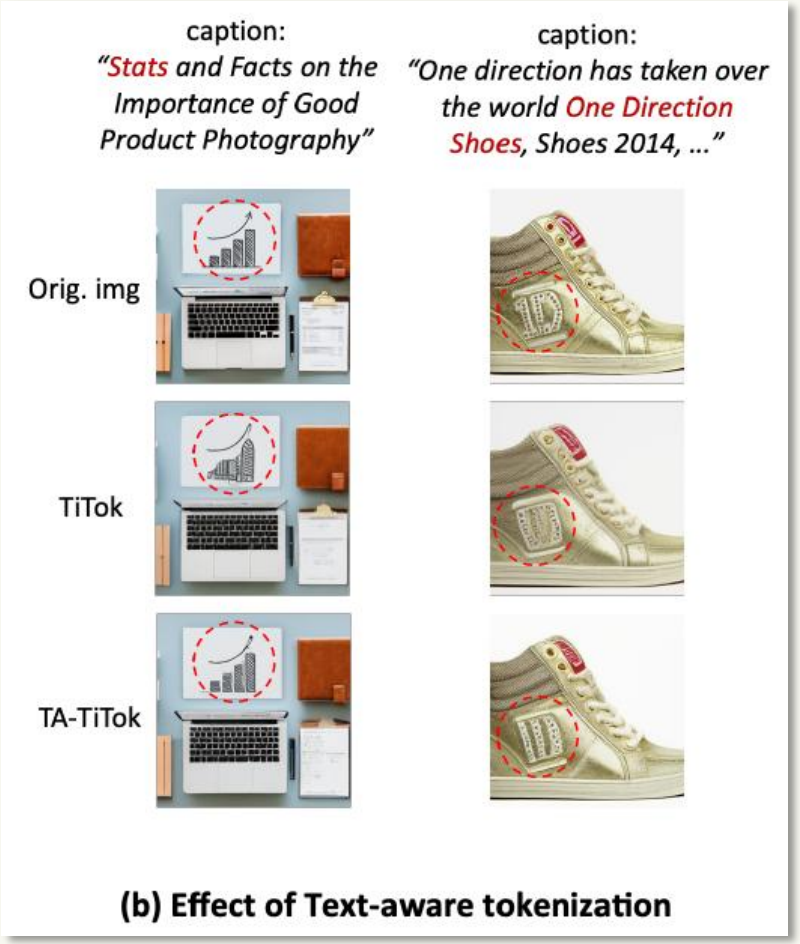
32 tokens

TiTok: 256×256 이미지를 32 discrete tokens로 표현

Source: TiTok / “An Image is Worth 32 Tokens for Reconstruction and Generation”, NeurIPS 2024

이미지 품질 개선 : Text 추가

토큰에 텍스트 의미를 주입하여 더 높은 품질의 이미지를 생성합니다.



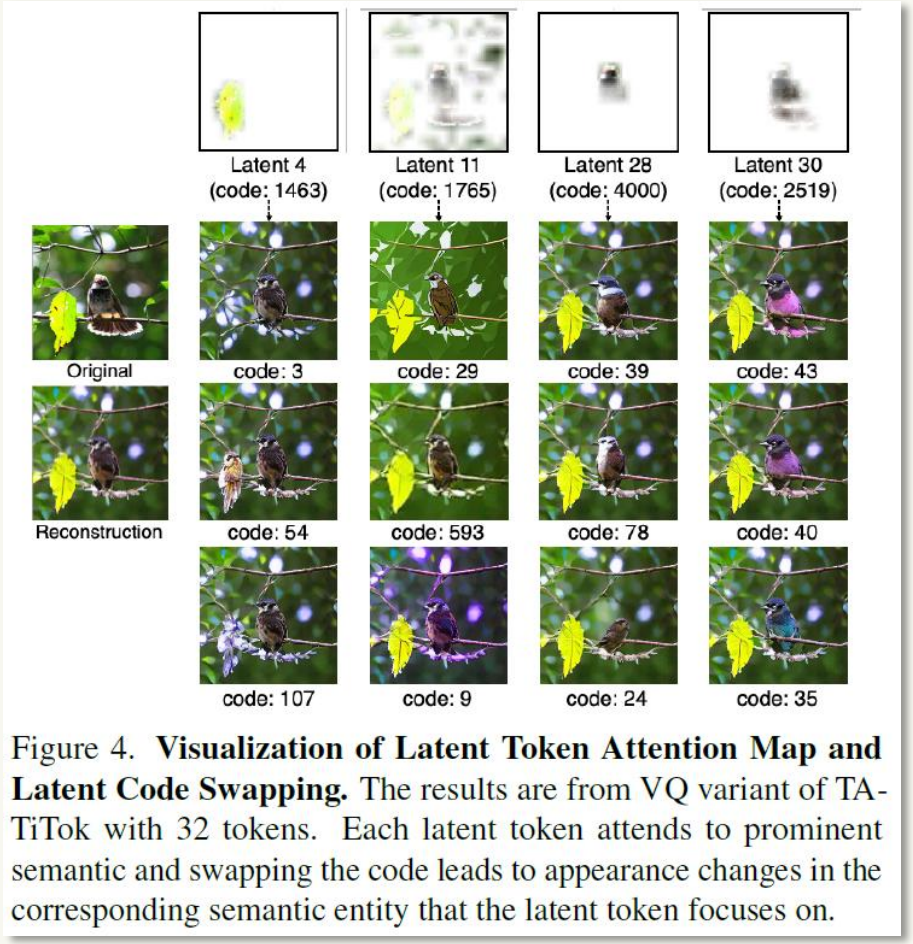


Figure 4. **Visualization of Latent Token Attention Map and Latent Code Swapping.** The results are from VQ variant of TA-TiTok with 32 tokens. Each latent token attends to prominent semantic and swapping the code leads to appearance changes in the corresponding semantic entity that the latent token focuses on.

Sources: TA-TiTok / "Democratizing T2I Masked Generative Models with Compact Text-Aware 1D Tokens", ICCV 2025

이미지 품질 개선: Codebook 키우기

codebook을 키워서 expressive하게 만들면서도, 동시에 codebook usage를 유지합니다.

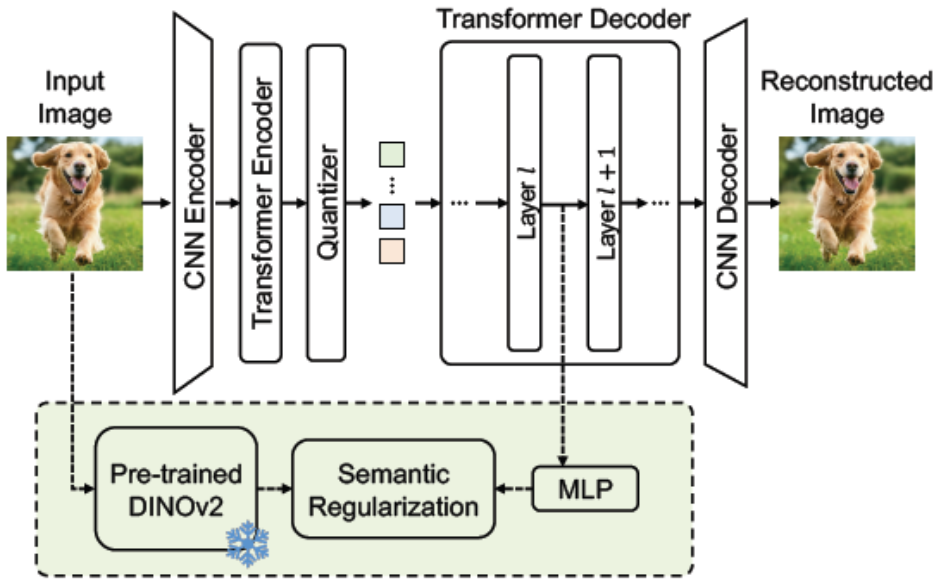


Figure 4. **GigaTok architecture and semantic regularization.** *Top:* We use a hybrid CNN-Transformer design for our visual tokenizer. The transformer layers are implemented with ViT for 2D tokenizer and Q-Former for 1D tokenizer. *Bottom:* We use a frozen DINOv2 [43] image encoder for semantic regularization.

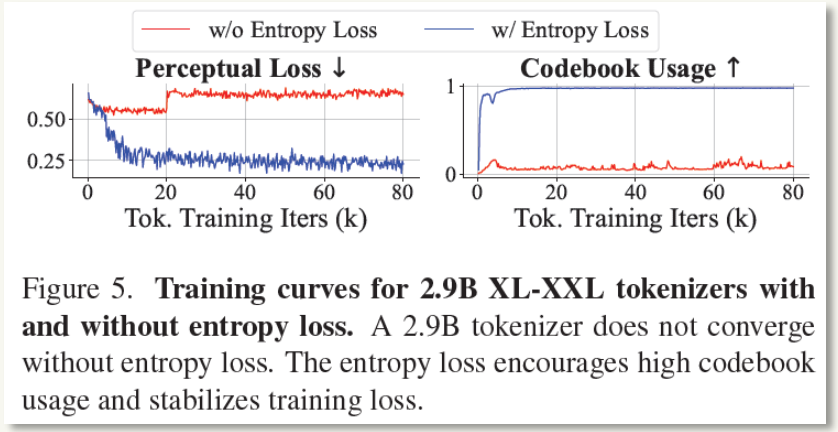


Figure 5. **Training curves for 2.9B XL-XXL tokenizers with and without entropy loss.** A 2.9B tokenizer does not converge without entropy loss. The entropy loss encourages high codebook usage and stabilizes training loss.

Figure 2. The 2.9B GigaTok achieves SOTA autoregressive image generation with a 1.4B AR model on ImageNet 256×256 resolution.

Sources: GigaTok / "Scaling Visual Tokenizers to 3B Parameters for AR Image Generation", ICCV 2025

Concept 추가

Sparse Auto Encoder를 이용하여 semantic 정보를 포함한 token을 얻습니다.

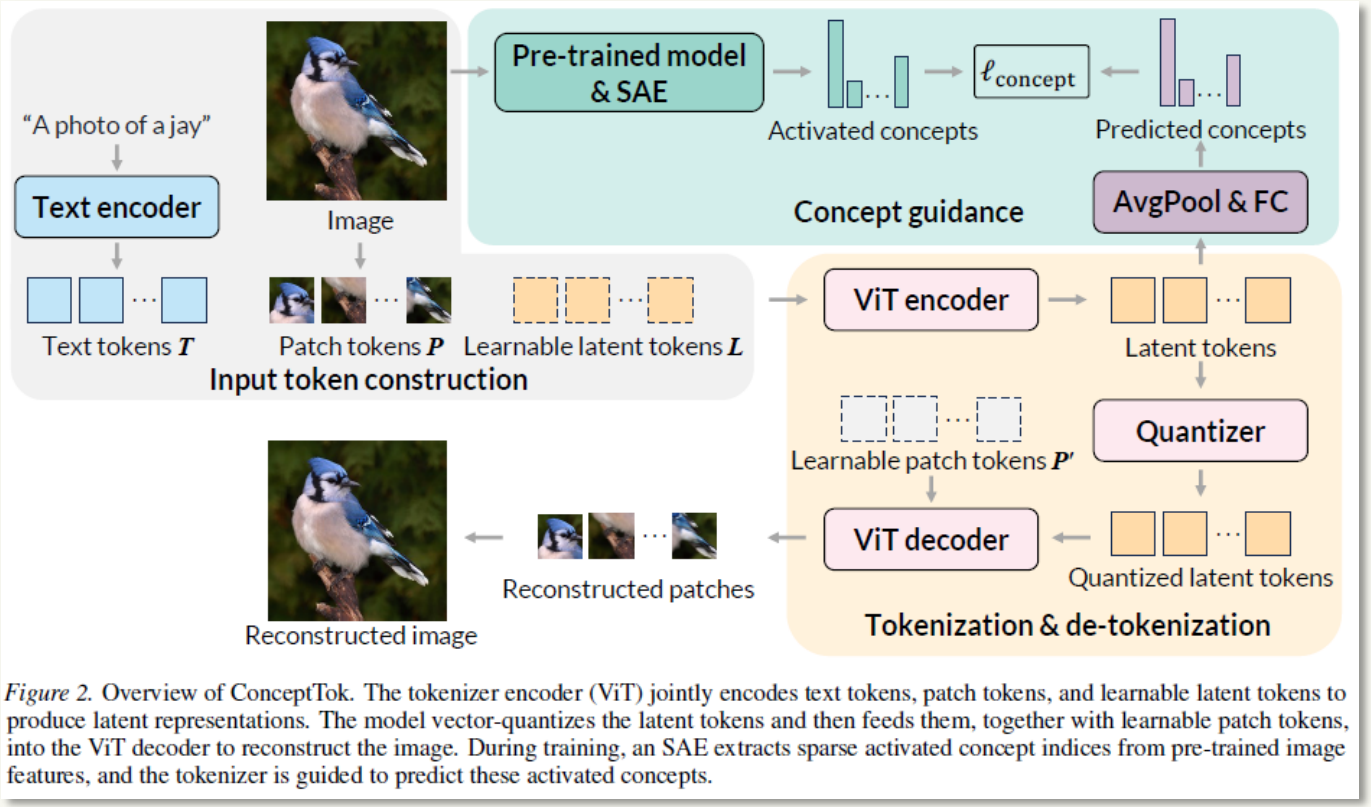


Figure 2. Overview of ConceptTok. The tokenizer encoder (ViT) jointly encodes text tokens, patch tokens, and learnable latent tokens to produce latent representations. The model vector-quantizes the latent tokens and then feeds them, together with learnable patch tokens, into the ViT decoder to reconstruct the image. During training, an SAE extracts sparse activated concept indices from pre-trained image features, and the tokenizer is guided to predict these activated concepts.

Figure 4. Qualitative comparisons of images generated from identical text prompts using different tokenizers. ConceptTok produces images with higher fidelity and better semantic alignment.

Sources: ConceptTok / "Concept-Guided Tokenization", ICML 2026

남아 있는 문제

Fine details

작은 글자, 얼굴, 손가락, texture등의 high-frequency 정보가 손실 될 수 있음

Recon vs. Gen

좋은 reconstruction 지표가 좋은 generation을 보장하지 않음

Efficiency vs. Fidelity

적은 token은 비용을 줄이지만 정보 손실을 키울 수 있음

Universal Vocabulary

이미지·비디오·언어·오디오 token들이 의미적으로 잘 정렬되는지