SiriTok - Semantically Rich 1D Tokenizer

Jason Zhao

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Background

DINO

DINO is a vision foundation model by Meta that trained to produce semantically meaningful image representations. Its contrastive learning strategy make it strong at capturing both the global (CLS) representation and per patch semantic understanding.

VAE

Variational Autoencoder (VAE) is a generative model that learn to represent real data (e.g. images) in a structured continuous latent space. In image generation models like Stable Diffusion, models generate images in this latent space instead of pixel space. Therefore, the structure and quality of the latent representation directly affect generative performance.

Diffusion Models

Diffusion models learn to transform a simple prior distribution (e.g., Gaussian noise) into a target data distribution (image latent distribution) through iterative denoising. If the latent space has stronger semantic structure, diffusion trajectories are more aligned with meaningful variations rather than low-level noise patterns.

1D Tokenization

Instead of preserving the 2D structure of the images like traditional techniques (e.g. VAE), it builds on a novel assumption that using a 1D space with K vectors is sufficient for encoding images, allowing a more dynamic arrangement of encodings on different parts of the image.

SiriTok

As discussed in the paper Titok, 1D tokenization of images has great potentials for fast image generation because of its nature of small amount of encoding tokens.

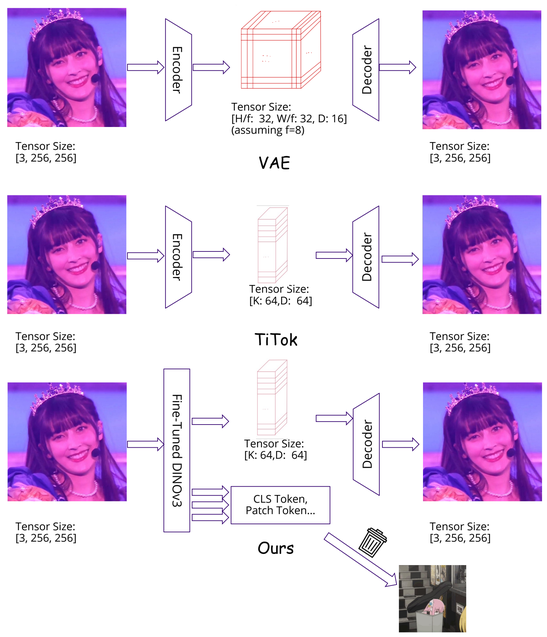
However, the authors of Titok found that while having a higher K (number of vectors to represent an image) would result in better reconstruction quality, it will also negatively affect the semantic information preserved by the encodings.

Recently, multiple works have tried to add semantic information into the VAE encoding space, such methods have shown great results of improving the learning speed of diffusion models.

Motivated by these prior works, in this project, we focus on introducing semantic information into the 1D tokens' world. We design novel objectives to force the encoder to preserve high level semantic information as well as the low-level image details.

We summarize our contribution as:

1. Novel semantic preservation loss for 1D encoders
2. Introducing semantic information into continuous 1D latent space



Method

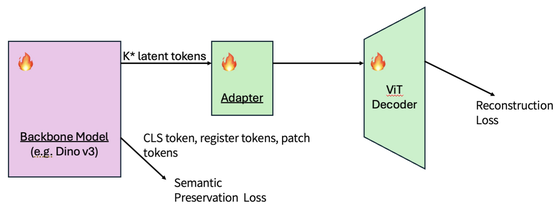
Our goal of this project is to train a 1D tokenizer with constraints of both detail reconstruction and semantic preservation. To achieve this goal, we fine tune a pretrained DINOv3 ViT as our encoder, and we design our to training objective to preserve the rich semantic prior of DINO. In detail, we use a combination of the following 5 losses:

Reconstruction Loss:

- **Pixel Level L1 loss:** Measures the mean absolute difference between reconstructed and ground-truth pixels
- **MS-SSIM loss:** Measures structural similarity across multiple scales (luminance, contrast, structure). Focuses on perceptual quality and texture consistency
- **LPIPS Loss:** Measures perceptual similarity from deep feature embeddings (e.g., VGG). Encourages semantic and texture-level similarity

Semantic Preservation Loss:

- **Patch Similarity Loss:** Measures the L2 difference of the original and fine tuned DINO's image patch encodings.
- **CLS Probing Loss:** Use an additional MLP + pooling layer with random dropouts, to reconstruct the DINO CLS from the latent tokens, measures the L1/L2 loss on the reconstructed token and the original DINO's CLS token to encourage semantic information to be evenly distributed on all latent representations.



Experiment

Dataset: We train our model on the train split of ImageNet, which contains 1.2M samples, and validate using its val split. We crop the images into square and resize them into 256 * 256 pixels using pillow's implementation of BICUBIC.

Model Architecture: For the backbone model, we chose `dinov3-vitb16-pretrain-lvd1689m` for a balance of quality and compute. The adapter is a simple linear module.

Training Settings: We train on an 8 A40 GPU node with DDP. Using an effective batch of 640. We train 2 copies with and without semantic preservation loss and kept every other parameters identical.

Linear Probing Evaluation: We employ a linear probing protocol following MAE and TiTok. The classifier consists of a single affine-free Batch Normalization layer followed by a linear projection to 1,000 ImageNet classes. This minimal architecture ensures that observed accuracy reflects the intrinsic quality of the token representations rather than any capacity in the classifier itself.

Results

Reconstruction Quality

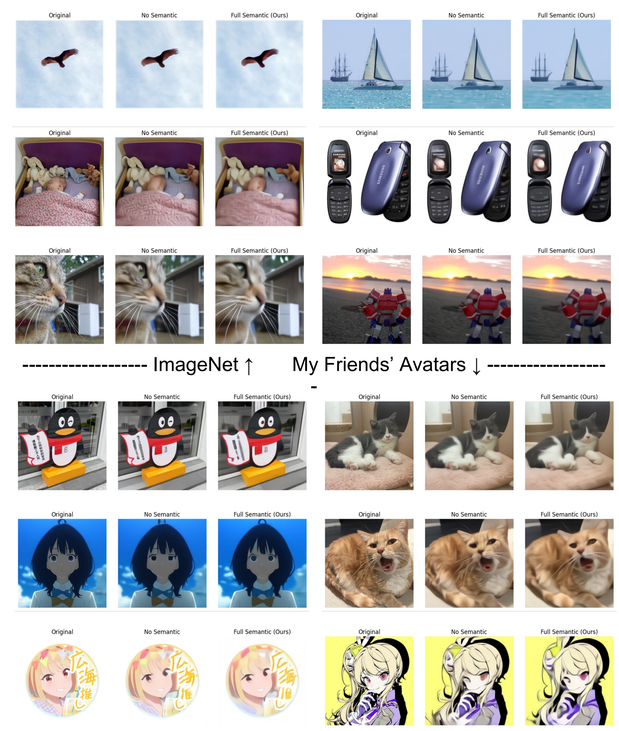


Figure 1. Qualitative comparison. We show comparable reconstruction quality for models trained with and without the semantic preservation loss. We show results from the ImageNet volition set on the *top* and out-of-distribution inputs on the *bottom*.

Semantic Preservation

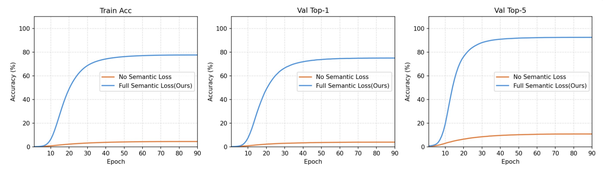


Figure 2. Linear probing results on the encoded latents of encoders trained with and without semantic preservation loss. Our semantic preservation loss helps greatly with preservation semantic information with no visible tradeoff for image quality.

Future Work

- Add GAN discriminators for better reconstruction results
- Test the diffusibility of our proposed method by training a diffusion model based on this latent space