“Principal Components” Enable A New Language of Images

Supplementary Material

Contents

Author Contribution Statement

X.W. and B.Z. conceived the study and guided its overall direction and planning. X.W. proposed the original idea of semantically meaningful decomposition for image tokenization. B.Z. developed the theoretical framework for nested CFG and the semantic spectrum coupling effect and conducted the initial feasibility experiments. X.W. further refined the model architecture and scaled the study to ImageNet. B.Z. led the initial draft writing, while X.W. designed the figures and plots. I.E., J.D., and X.Q. provided valuable feedback on the manuscript. All authors contributed critical feedback, shaping the research, analysis, and final manuscript.

Limitations and Broader Impacts

Our tokenizer contributes to structured visual representation learning, which may benefit image compression, retrieval, and generation. However, like other generative models, it could also be misused for deepfake creation, misinformation, or automated content manipulation. Ensuring responsible use and implementing safeguards remains an important consideration for future research. SEMANTICIST also presents several limitations, for example, we employ a diffusion-based decoder, but alternative generative models like flow matching or consistency models could potentially improve efficiency. Additionally, our framework enforces a PCA-like structure, further refinements, such as adaptive tokenization or hierarchical models, could enhance flexibility.

A. Proof for PCA-like structure

The conditional denoising diffusion model is using a neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}, t)$ to approximated the score function $\nabla_{\boldsymbol{x}_t} \ln q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ which guides the transition from a noised image $\boldsymbol{x}_t$ to the clean image $\boldsymbol{x}_0$. For the conditional diffusion decoder in SEMANTICIST, the score function can be decomposed as:

$$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_1, \dots, \boldsymbol{z}_k) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \emptyset) + \sum_{i=1}^{k} \gamma_i \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_i)\,,$$

where $\emptyset$ is the null condition, γ_i is the guidance scale, and $\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_i) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_1, \dots, \boldsymbol{z}_i) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_1, \dots, \boldsymbol{z}_{i-1})$ represents the increment contribution of the concept token condition $\boldsymbol{z}_i$ to the score function. Thus, we can rewrite the diffusion training objective with k conditions with the following:

$$\mathcal{L}_k = \mathbb{E}\left[\left\|\epsilon - \left(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \emptyset) + \sum_{i=1}^{k} \gamma_i \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_i)\right)\right\|^2\right].$$

Orthogonality between contribution of concept tokens. At the optimal convergence, the gradient of $\mathcal{L}_k$ w.r.t $\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_i)$ is zero, thus give us:

$$\begin{aligned}\frac{\partial \mathcal{L}_k}{\partial \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_i)} &= \mathbb{E}\Big[\big(\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \emptyset) - \sum_{j=1}^{k} \gamma_j \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_j)\big)\gamma_i\Big] \\ &= 0\,.\end{aligned}$$

Since model is at convergence, the residual term $\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \emptyset) - \sum_{j=1}^{k} \gamma_j \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_j)$ can not be further reduced by making further changes to the adjustment from the i-th concept token $\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_j)$. In other words, the residual term and all active conditions $\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_j)$ are orthogonal to each other. Next, we can use induction to prove that at convergence, all $\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_j)$ terms are orthogonal to each other similar to PCA. For the case of $k = 1$, we only use one concept token to condition the model, thus we can have:

$$\mathbb{E}\left[\left(\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \emptyset) - \gamma_1 \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_1)\right) \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_1)\right] = 0\,.$$

For the case of $k = 2$, for $(i = 1, 2)$, we have:

$$\mathbb{E}\left[\Big(\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \emptyset) - \sum_{j=1}^{2} \gamma_j \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_j)\Big) \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_i)\right] = 0\,.$$

By substituting the $k = 1$ case into this, it can be seen that $\mathbb{E}\left[\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_1)^\top \Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{z}_2)\right] = 0$. Assuming this orthogonality holds for the first $k-1$ concept tokens: $\mathbb{E}\left[\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_i)^\top\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_j)\right] = 0 \quad \forall i,j < k, i \neq j$. Then for $i < k$, by substituting

$$\epsilon - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\emptyset) = \sum_{j=1}^{k-1}\gamma_j\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_j) + \gamma_k\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_k)\,,$$

we can have:

$$\mathbb{E}\left[\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_i)^\top\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_k)\right] = 0\,.$$

Thus, the orthogonality propagates to all pairs $(\boldsymbol{z}_i,\boldsymbol{z}_k)$ for $i < k$. By induction, we have orthogonality between all pairs of concept tokens.

Variance Explained Hierarchy. Assuming the true noise ϵ can be reconstructed using the conditional model, we have:

$$\epsilon \approx \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\emptyset) + \sum_{i=1}^{k}\gamma_i\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_i) + \text{residual}\,.$$

Given the orthogonality of $\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\boldsymbol{z}_i)$ we have proven earlier, the total variance can be decomposed as:

$$\text{Var}(\epsilon) = \sum_{i=1}^{k}\text{Var}(\gamma_i\Delta\boldsymbol{\epsilon}(\boldsymbol{x}_t,\boldsymbol{z_i})) + \text{Var}(\text{residual})\,.$$

Let $\lambda_i = \text{Var}(\gamma_i\Delta\boldsymbol{\epsilon}_\theta(\boldsymbol{x_t},\boldsymbol{z}_i))$, representing the variance explained by concept token condition $\boldsymbol{z}_i$. Our dropout design would have the training objective forces:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k\,,$$

as each concept token $\boldsymbol{z}_i$ is trained to explain the maximal residual variance after accounting for concept tokens $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_{i-1}$.

Thus, combining the orthogonality and the variance decay, SEMANTICIST provably grounds the emergence of a **PCA-like hierarchical structure** in the learned concept tokens. Providing a simple, effective, and explainable architecture for visual tokenization.

B. Additional Related Work

B.1. Concurrent Related Work

Concurrent work [1] introduces a 1D tokenizer that focuses on adaptive-length tokenization by resampling sequences of 1D tokens from pre-trained 2D VAE tokens. In contrast, our encoder builds on raw RGB images. More importantly, our approach is motivated by a fundamentally different objective — reintroducing a PCA-like structure into visual tokenization to enforce a structured, hierarchical latent representation. Furthermore, our tokenizer is continuous rather than discrete, setting it apart from [1] and allowing it to better capture the variance-decaying properties inherent to PCA. Additionally, we identify and resolve the semantic-spectrum coupling effect, a key limitation in existing visual tokenization methods that have not been previously addressed.

B.2. Related Work on Human Perception

Human perception of visual stimuli has been shown to follow the global precedence effect [34], where the global information of the scene is processed before the local information. In [15], controlled experiments of presentation time on human perception of visual scenes have further confirmed the global precedence effect, where less information (presentation time) is needed to access the non-semantic, sensory-related information of the scene compared to the semantically meaningful, object- or scene-related information. Similar results have been reported in [3], where sensory attributes are more likely to be processed when the scene is blurred. Moreover, [35] has suggested that reliable structural information can be quickly extracted based on coarse spatial scale information. These results suggest that human perception of visual stimuli is hierarchical, where the global information of the scene is processed before the local information. As we have shown in the main paper, SEMANTICIST can naturally emerge a similar hierarchical structure in the token sequence, where the first few tokens encode the global information of the scene and the following tokens encode the local information of the scene. This hierarchical structure is provably PCA-like, similar to the hierarchical nature of human perception of visual stimuli.

B.3. Related Work on Diffusion-Based Tokenizers

The usage of a diffusion-based decoder has been explored by several works [7, 17, 65]. Zhao et al. [65] proposed the usage of a diffusion-based decoder as a paradigm shift from single-step reconstruction of previous tokenizers to the diffusion-based iterative refinement process. Chen et al. [7] further scale this idea on more modern DiT [37] architecture and describe the scaling law for such diffusion-based tokenizers. Ge et al. [17] applied this idea to a video tokenizer, enabling better reconstruction and understanding of video content. However, these previous works overlook the benefit of the diffusion-based decoder in that it can disentangle the semantic content from the spectral information. Additionally, these works still apply the 2D grid-based structure for encoding the image without considering the latent structure of the token space.

C. Additional Implementation Details

C.1. Semanticist Autoencoder

Model architecture. As shown in Fig. 3, the SEMANTICIST tokenizer follows the diffusion autoencoder [28, 38] paradigm: a visual encoder takes RGB images as input and encodes them into latent embeddings to condition a diffusion model for reconstruction. In our case, the visual encoder is a ViT-B/16 [12] with a sequence of concept tokens concatenated with image patches as input. The concept tokens have full attention with patch tokens, but are causal to each other. Before being fed to the decoder, the concept tokens also go through a linear projector, and are then normalized by their mean and variance. To stabilize training, we also apply drop path with a probability of 0.1 to the ViT. For the DiT decoder, we concatenate the patch tokens (condition) with noisy patches as input, and the timesteps are still incorporated via AdaLN following common practice [37].

Nested classifier-free guidance (CFG). For the DiT decoder, we randomly initialize k (number of concept tokens) learnable null-conditioning tokens. During each training iteration, we uniformly sample a concept token index k', and corresponding null tokens replace all tokens with larger indices. To facilitate the learning of the encoder, we do not enable nested CFG in the first 50 training epochs. During inference, CFG can be applied to concept tokens independently following the standard practice [37].

Training. We follow [28] for training details of the tokenizer. Specifically, the model is trained using the AdamW [31] optimizer on ImageNet [8] for 400 epochs with a batch size of 2048. The base learning rate is 2.5e-5, which is scaled by $lr = lr_{\text{base}} \times \text{batch size}/256$. The learning rate is also warmed up linearly during the first 100 epochs, and then gradually decayed following the cosine schedule. No weight decay is applied, and β_1 and β_2 of AdamW are set to 0.9 and 0.95. During training, the image is resized so that the smaller side is of length 256, and then randomly flipped and cropped to 256×256. We also apply a gradient clipping of 3.0 to stabilize training. The parameters of the model are maintained using exponential moving average (EMA) with a momentum of 0.999.

Inference. Because of the nature of the PCA structure, it is possible to obtain reasonable reconstruction results with only the first few concept tokens. In implementation, we achieve this by padding missing tokens with their corresponding null conditioning tokens and then feeding the full sequence to the DiT decoder.

C.2. Autoregressive Modeling

Model architecture. The ϵLlamaGen roughly follows the LlamaGen architecture with the only change of using a diffusion MLP as the prediction head instead of a softmax head. To perform the classifier-free-guidance, we use one `[CLS]` token to guide the generation process of ϵLlamaGen. As certain configurations of SEMANTICIST can yield high-dimensional tokens, we made a few adjustments to the model architecture of ϵLlamaGen to allow it to learn with high-dimensional tokens. Specifically, we use a 12-layer MLP with each layer having 1536 hidden neurons as the prediction head and use the stochastic interpolant formulation [32] to train the diffusion MLP. The classifier-free guidance is also slightly modified: we concatenate the `[CLS]` token with the input to the diffusion MLP along the feature axis and then project back to the original feature dimension to feed into the diffusion MLP. These changes allow us to train auto-regressive models on high-dimensional (e.g., 256-dimensional) tokens with improved stability compared to the original version proposed in [29]. However, we expect future research to drastically simplify this model architecture.

Training. The ϵLlamaGen is trained for 400 epochs with cached latents generated by pretrained SEMANTICIST on the ImageNet dataset with TenCrop and random horizontal flipping augmentations. We use a batch size of 2048, and apply a 100-epoch warmup for the base learning rate of 1e-4, which is scaled similarly as the SEMANTICIST w.r.t. the batch size. After warmup, the learning rate is fixed. Weight decay of 0.05 and gradient clipping of 1.0 are applied. In our experiments, we find that later concept tokens have diminishing returns or are even harmful for ϵLlamaGen, thus only train ϵLlamaGen with the first few tokens. Specifically, the ϵLlamaGen-L model is trained with 32 concept tokens.

Inference. In the inference stage, we use the same linear classifier-free guidance schedule as MAR [29] and MUSE [5]. The schedule tunes down the guidance scale of small-indexed tokens to improve the diversity of generated samples, thus being more friendly for gFID. When reporting gFID, we disable CFG for SEMANTICIST's DiT decoder, tune the guidance scale of the autoregressive model, and report the best performance.

C.3. Linear Probing

We utilized the `sklearn` library to perform the linear probing experiments, the encoder weights are frozen, and we encode each image to its token representation. The linear classifier is learned on the token space without applying any data augmentation.

D. Additional Experiment Results

D.1. Human Perception Test

We are interested in understanding whether the tokens learned by SEMANTICIST follow a human-like perception effect, namely the global precedence effect [34] where the global shape and semantics are picked up within a very short period of exposure. Thus, we designed a human perception test to evaluate whether SEMANTICIST generates tokens that closely follow human perception. Specifically, we generate

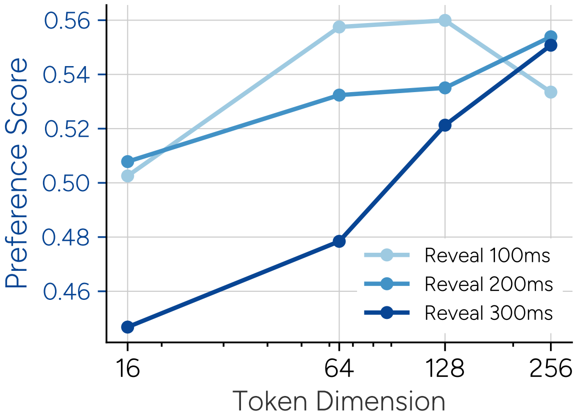

Figure 8. The preference score from the human perception test, all models and test configurations obtained a score close to 0.5, indicating SEMANTICIST can encode images as effectively as human language does.

images by only reconstructing from the first two tokens from SEMANTICIST. Distractor images are also generated by first captioning the image with Qwen2.5VL [2] and then generate the image with a stable-diffusion model [45]. Following the setup of [15], we only reveal the generated images and the distractors by a very short reveal time, and then ask the participants to choose which images more closely align with the original image. For evaluation, we give the participant's preference to distractor image zero points, the preference to the generated image one point, and in the case of a tie, we give 0.5 points. Fig. 8 presents the averaged preference score with different token dimensions and reveal time. SEMANTICIST is able to obtain a score close to 0.5 under all cases, indicating that SEMANTICIST can encode the image's global semantic content close to how state-of-the-art vision language models [2] encode the image in language space. A web-based human perception test interface is provided along with this appendix.

D.2. Zero-Shot CLIP on Reconstructed Images

We also study the property of the SEMANTICIST latent space by reconstructing from it. Fig. 9 demonstrate the zero-shot accuracy of a pretrained CLIP [40] model on the imagenet validation set reconstructed by SEMANTICIST. For all model variants, the zero-shot performance improves with the number of tokens, with models using more dimensions per token achieving better performance with a smaller number of tokens, indicating that with more dimensions, SEMANTICIST is able to learn the semantic content with fewer tokens. Fig. 6 provides the rFID score on the ImageNet validation set with a varying number of tokens, similar conclusions can be drawn. Additionally, Fig. 6 also provides the scaling behavior of SEMANTICIST, we can observe that SEMANTICIST not only enjoys a structured latent space, but also demonstrates a promising scaling.

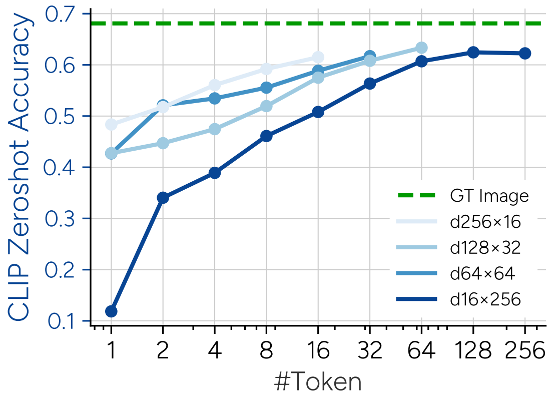

Figure 9. CLIP zero-shot accuracy on reconstructed images.

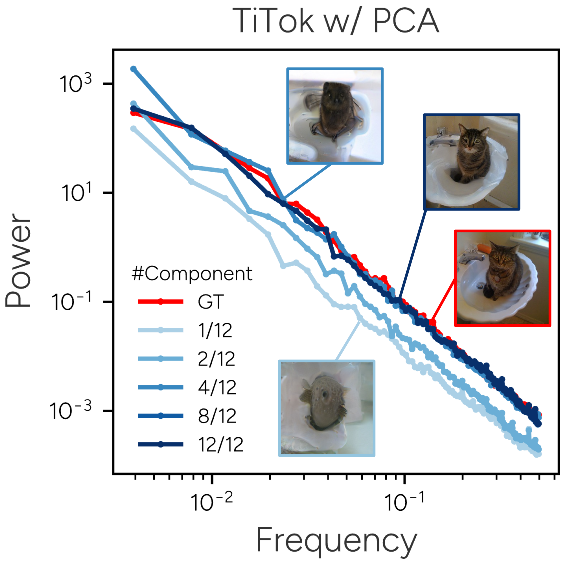

Figure 10. Frequency-power spectra of TiTok decomposed with PCA at feature dimensions. The learning of semantic contents and spectral information is coupled.

D.3. Semantic Spectrum Coupling Effect Results

In Fig. 10, we present the power frequency plot of performing PCA to decompose the latent token space of TiTok [62]. A similar effect as the PCA decomposition on VQ-VAE [50] and the first k token decomposition on TiTok [62] is observed. This result further demonstrates that the latent space of TiTok [62] entangles the semantic contents and the spectral information.

D.4. Additional Ablation Study

In Fig. 12, we show the results of SEMANTICIST with d64×64 tokens trained with or without REPA [63] evaluated by reconstruction FID on ImageNet 50K validation set. Despite the performance with full tokens being similar, adding REPA significantly improves the contribution of each (especially the first few) tokens. This naturally fits our need for PCA-like structure and is thus adopted as the default.

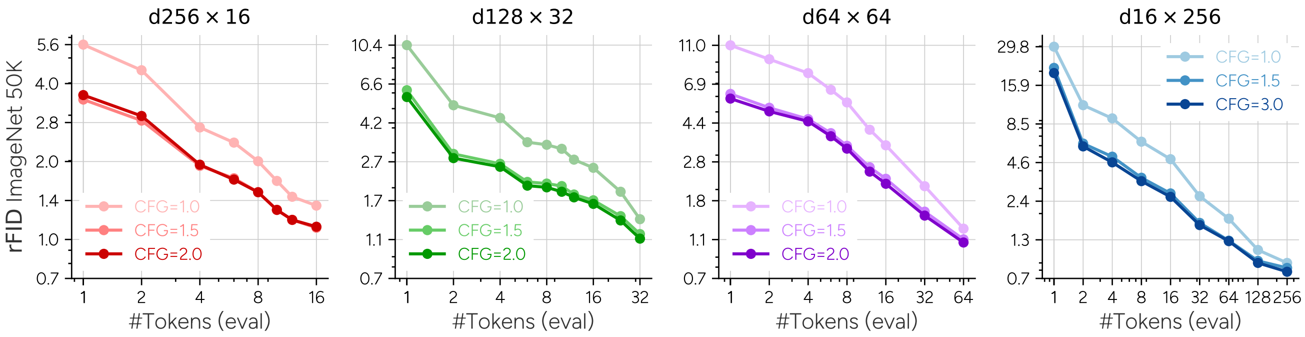

Figure 11. Reconstruction performance of different encoder configurations on ImageNet val 50K benchmark. A larger number of lower-dimensional tokens is more friendly for reconstruction tasks.

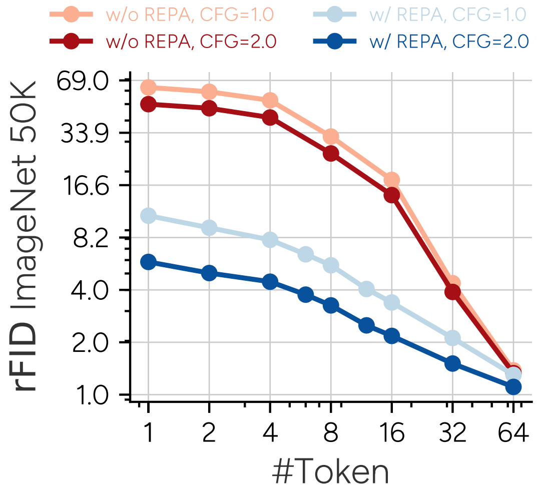

Figure 12. Ablation on the use of REPA (with d64×64 concept tokens, DiT-L/2 decoder, see qualitative results in Fig. 16). REPA improves the information density in preceding tokens.

We also compared the reconstruction performance of different concept token dimensions. We fix the product between the number of tokens and the dimension per token to be 4096, and investigate 256-dimensional (d256×16), 128-dimensional (d128×32), 64-dimensional (d64×64), and 16-dimensional (d16×256) tokens. As shown in Fig. 11, all configurations can learn ordered representations, with higher-dimensional ones containing more information per token. However, lower-dimensional tokens are more friendly for reconstruction tasks as they achieve better rFID.

D.5. Qualitative Results

In Fig. 13, reconstruction results from using different numbers of token dimensions are presented. As the dimension for one token becomes large, more semantic content can be encoded into it, thus allowing SEMANTICIST to generate faithful reconstructions of the original image.

In Fig. 14, the reconstructed results for different scaled DiT decoders are presented. These models are trained with the same dimension for the tokens that are 16-dimensional. We can see that as the model scales up, the reconstructed images with fewer tokens become more and more realistic and appealing.

Fig. 15 shows the reconstruction of the same SEMANTICIST tokenizer with different CFG guidance scales at inference time (CFG=1.0 indicates not applying CFG). It can be seen that the guidance scale has a very strong correlation with the aesthetics of generated images.

Fig. 16 presents qualitative results with or without the usage of REPA [63]. It is clear that the usage of REPA did not visually improve the final reconstruction by much, yet with fewer tokens, the model with REPA demonstrates more faithful semantic details with the original image.

Fig. 17 demonstrates the reconstruction results of more randomly sampled images, and Fig. 18 illustrates more intermediate results of auto-regressive image generation.

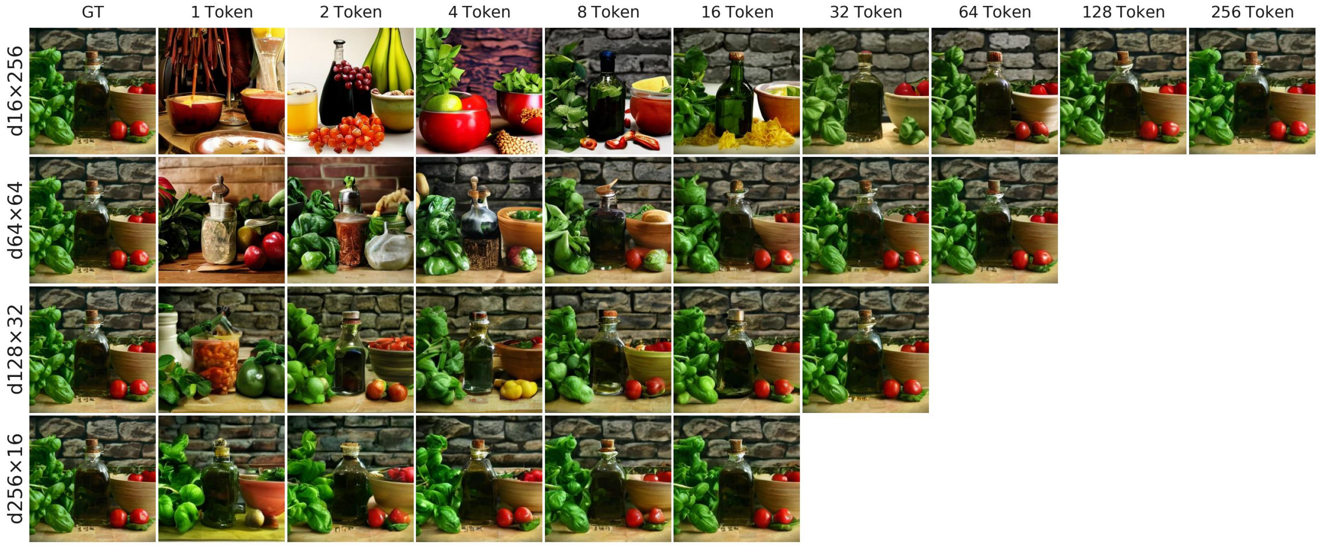

Figure 13. Qualitative results of different token dimensions. Higher-dimensional tokens encode more information, and lower-dimensional tokens achieve clearer semantic decoupling and better reconstruction.

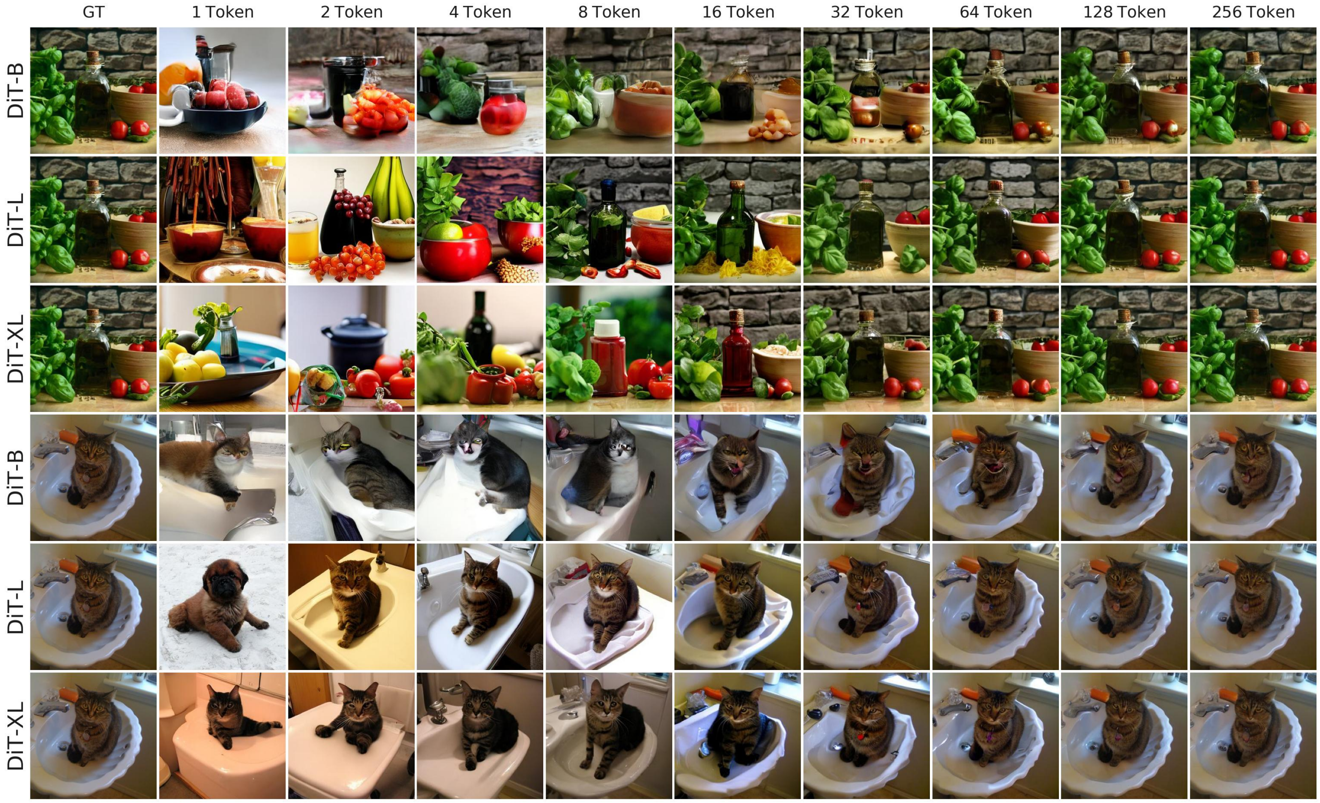

Figure 14. Qualitative results of different DiT decoder scales (DiT-B/2, DiT-L/2, and DiT-XL/2) with d16×256 tokens. The quality of images generated with fewer tokens improves consistently as the decoder scales up.

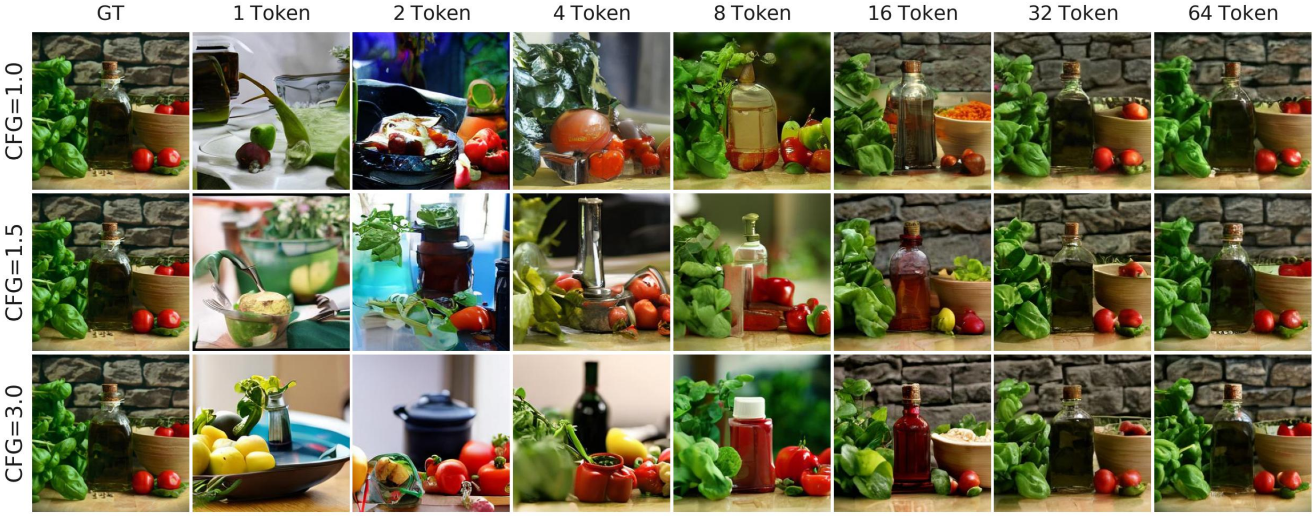

Figure 15. Qualitative results of different CFG guidance scales for DiT decoder, which clearly controls image aesthetics.

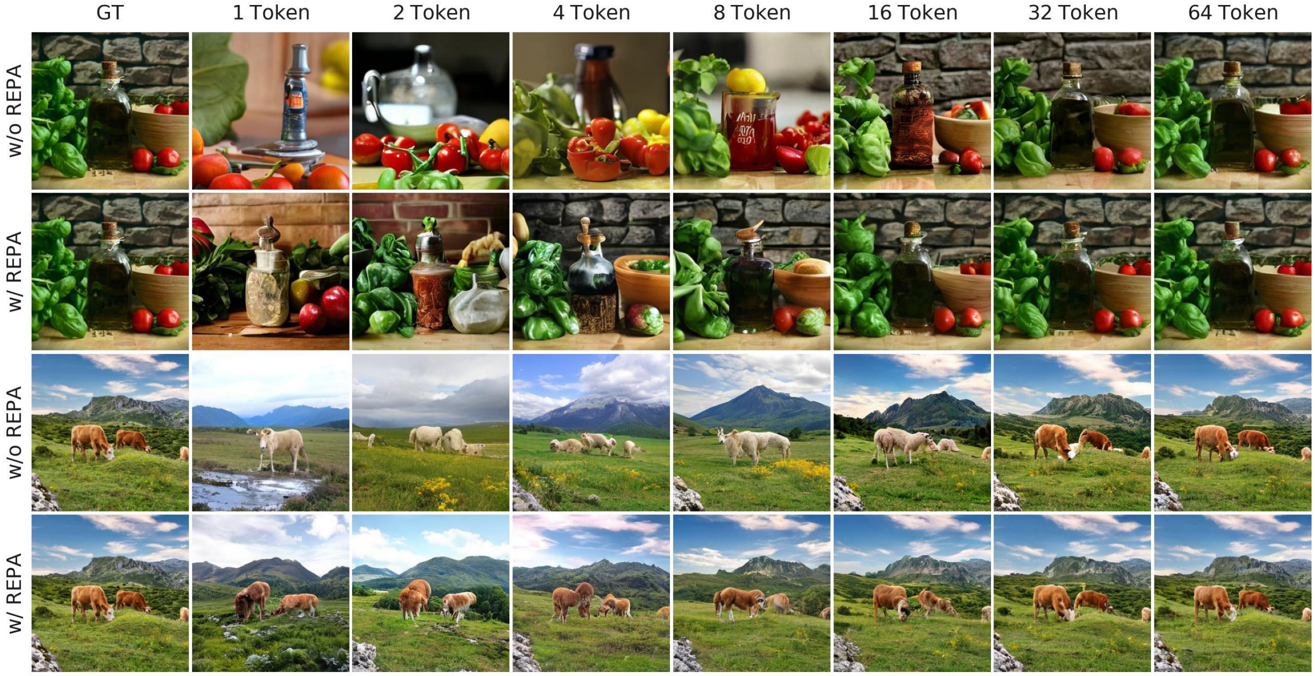

Figure 16. Qualitative results on effects of REPA (with d64×64 concept tokens). Instead of improving final reconstruction much, the benefit of REPA is mainly attributed to more faithful semantics in intermediate results.

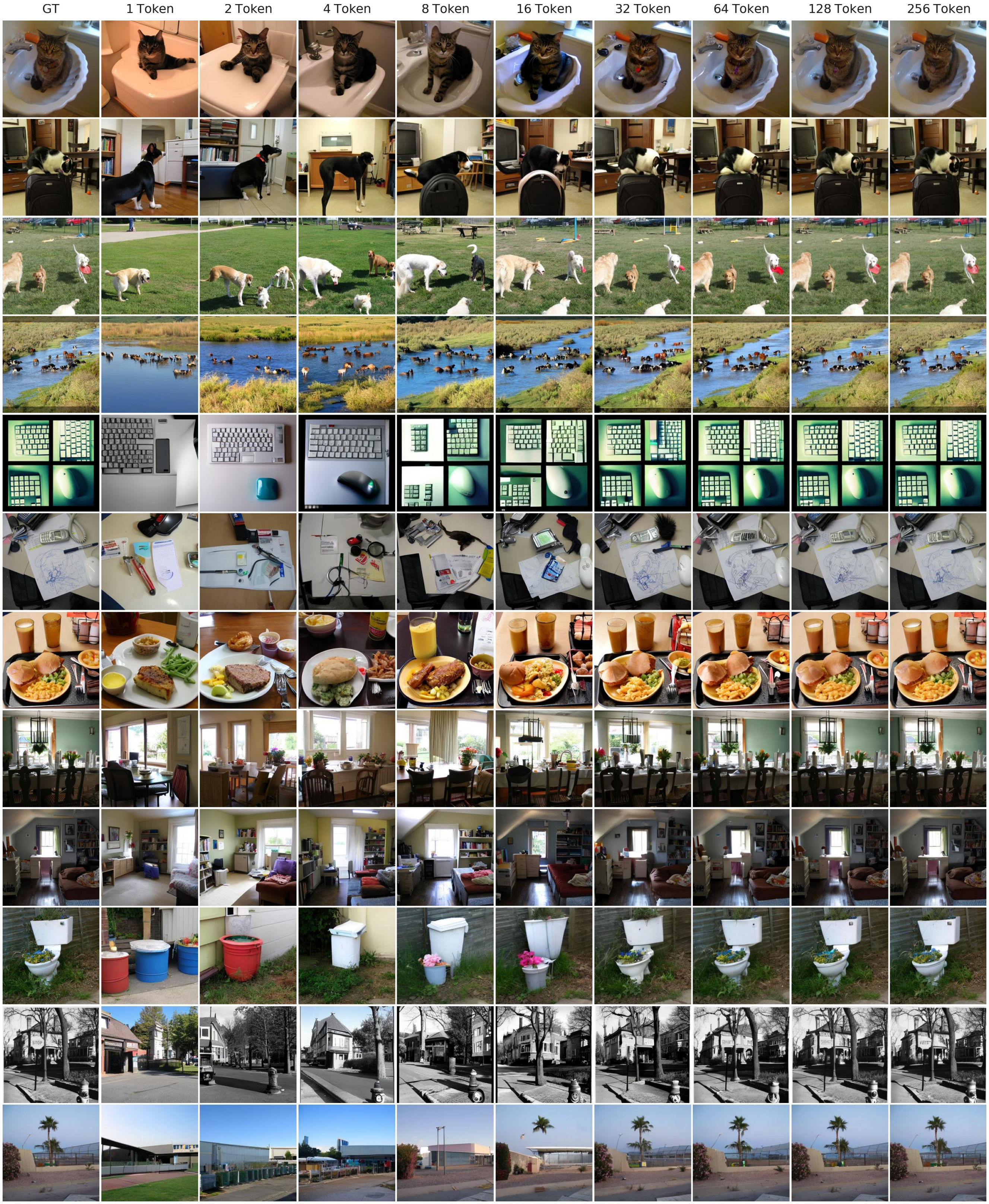

Figure 17. More reconstruction results of SEMANTICIST autoencoder (with d16×256 concepts tokens and DiT-XL/2 decoder).

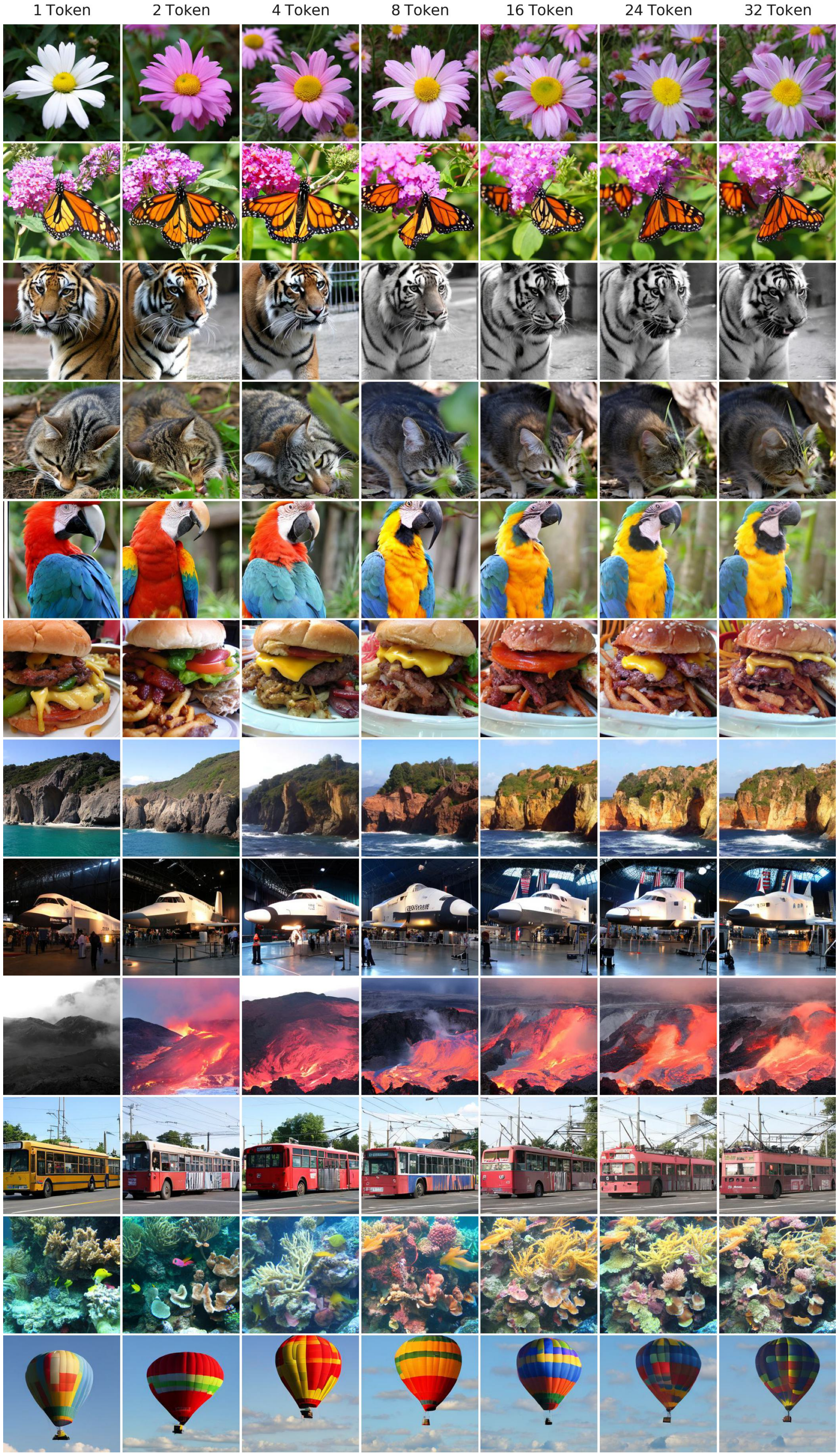

Figure 18. More visualization of intermediate results of auto-regressive image generation.