Supplementary Material: Adapting Self-Supervised Representations as a Latent Space for Efficient Generation

A Implementation Details

Generative Decoder Our generative decoder is implemented as a DiT-XL/2 (Peebles & Xie, 2023) and trained for one million steps with a learning rate of 10^{-4} using the AdamW (Loshchilov & Hutter, 2019) optimizer, a linear warm-up of 2000 steps and a global batch size of 512 on 8 H100 GPUs. Our implementation uses RoPE (Su et al., 2023; Crowson et al., 2024), RMSNorm (Zhang & Sennrich, 2019) and SwiGLU (Shazeer, 2020) activation functions, as we find that these modifications improve the stability and performance of our generative decoder. We concatenate the SSL embedding to the decoder patch tokens and apply full self-attention over all tokens.

MLP Mixer We adopt a standard MLP-Mixer (Tolstikhin et al., 2021) architecture, where all conditioning information: CLIP text embeddings for text-to-image (T2I) generation and class tokens for class-conditional image generation is concatenated with the noisy image token and passed through the model. Our implementation follows the configuration provided by the *lucidrains* [1] GitHub repository, with a hidden dimension of 1280, a depth of 28 layers, an expansion factor of 4 for the channel MLP, and 2 for the token MLP.

B Additional results

Qualitative examples per token type. As discussed in the main paper, DINOv2 (Oquab et al., 2024) offers two different types of tokens (besides patch tokens). First, the standard `[cls]` token and additionally a set of register tokens (Darcet et al., 2024). In Figure S3 we provide a qualitative comparison of the differences in outcome between these two token types. We keep the SSL backbone frozen and only train our generative decoder. We can observe that the `[reg]` token contains more knowledge about appearance, location, and object orientation compared to the `[cls]` token. However, none of the approaches gives proper pixel-wise reconstructions, again highlighting the need to integrate further information from the SSL encoder.

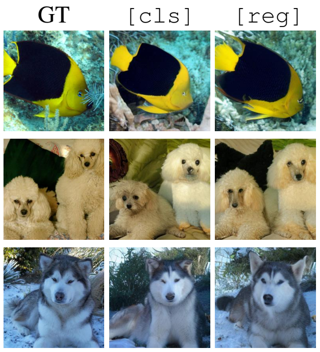


Figure S3: **`[cls]` vs `[reg]`** qualitative comparison.

Performance vs test-time compute. Figure S2a shows the number of function evaluations (NFE) vs reconstruction FID (rFID) on the ImageNet (Deng et al., 2009) validation dataset. Performance improves with increasing number of function evaluations (NFE), but saturates around 15–20. We hypothesize that the strong conditioning signal from the generative decoder reduces the need for additional refinement steps. Additionally, Figure S2b shows the generation performance depending on the number of sampling steps. We see a continuous improvement with more NFE.

Token type DINOv2 (Oquab et al., 2024) provides access to both a `[cls]` token and a set of register tokens. We compare their usefulness as latent representations for our generative decoder in Table 5. Using a frozen `[cls]` token results in strong reconstruction FID, indicating good semantic alignment, but yields low pixel-level scores such as PSNR and SSIM. In contrast, the register token captures more fine-grained visual details, improving pixel-wise reconstruction quality. This suggests that while the `[cls]` token emphasizes semantic content, the register token retains more low-level and regional information.

Table 5: **Ablation of token type.** Conditioning on DINOv2's (Oquab et al., 2024) register tokens improves pixel-wise metrics, indicating stronger local information.

Token	rFID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
`[reg]`	14.90	**12.85**	**29.07**	**0.52**
`[cls]`	**14.13**	12.59	28.41	0.54

[1] https://github.com/lucidrains/mlp-mixer-pytorch

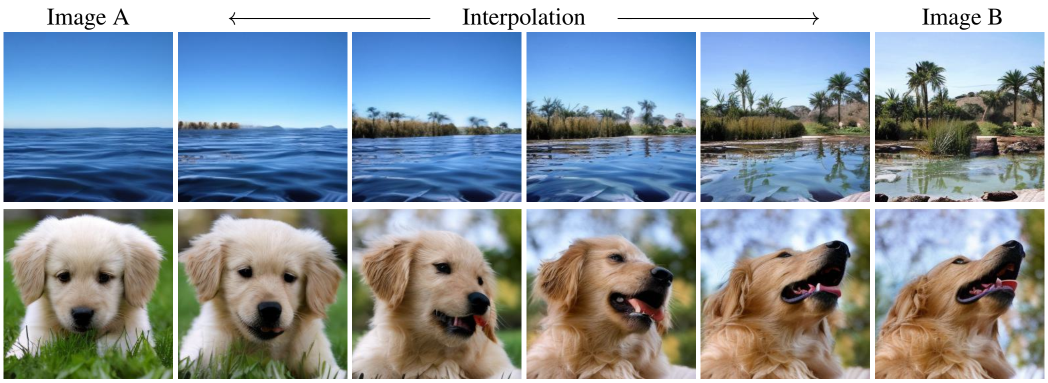
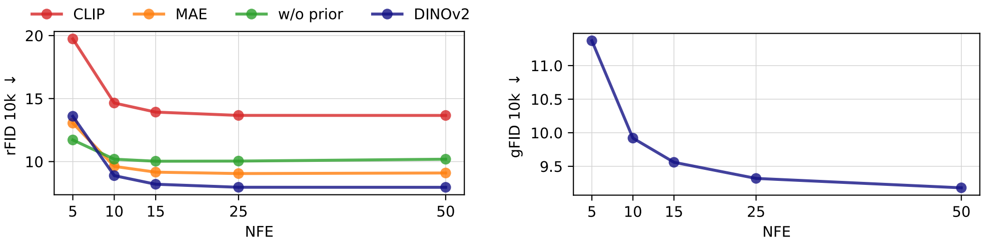


Figure S1: **More single token latent space interpolation results.** We observe smooth transitions not only in semantic content but also in object spatial configuration, and especially in object rotation (see dog).



(a) reconstruction

(b) generation

Figure S2: **Effect of inference steps** for reconstruction and generation.

More qualitative samples We provide additional qualitative results to further illustrate the capabilities of our model: text-conditional generations are shown in Figure S9, and uncurated class-conditional ImageNet generations in Figure S11.

Additional Tokens Some SSL encoders, such as DINOv2, provide additional global register tokens beyond the `[cls]` token. Incorporating these tokens increases latent capacity and improves reconstruction quality (see Figure S5, left). However, these tokens are typically unregularized and therefore do not inherit the favorable semantic and well-structured properties that our approach relies on. This also shows in the worse generative performance (gFID in Figure S5, right). Moreover, using multiple global tokens requires SSL models that have additional non-spatial tokens, as there is no straightforward way to repurpose spatial tokens into meaningful global representations. Given our focus on efficiency and structured single-token representations, we therefore restrict our method to the `[cls]` token, though multi-token extensions remain an interesting direction for future work.

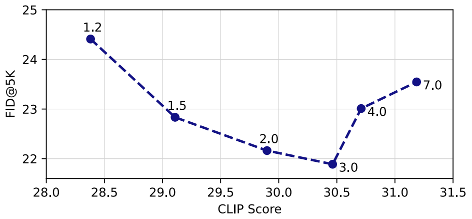


Figure S4: CFG effects on CLIP score and FID on the COCO (Lin et al., 2014) validation set. As commonly observed, CLIP score increases with larger CFG scales, while FID improves only within a moderate range before rising again.

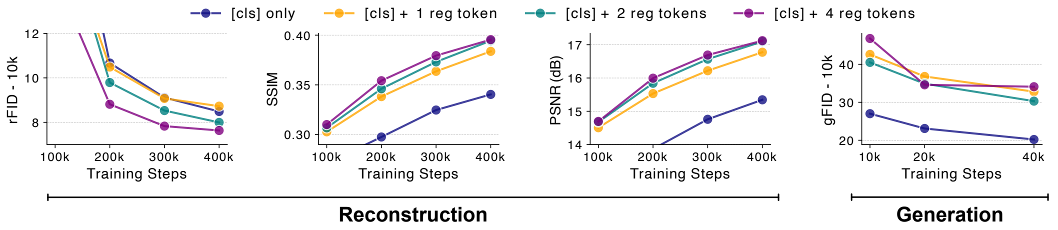


Figure S5: **Token number ablation.** Increasing the number of tokens in form of additional register tokens from DINOv2 Darcet et al. (2024) improves reconstruction quality. However, since register tokens do not have the favorable SSL properties, their space is not amenable for generation.

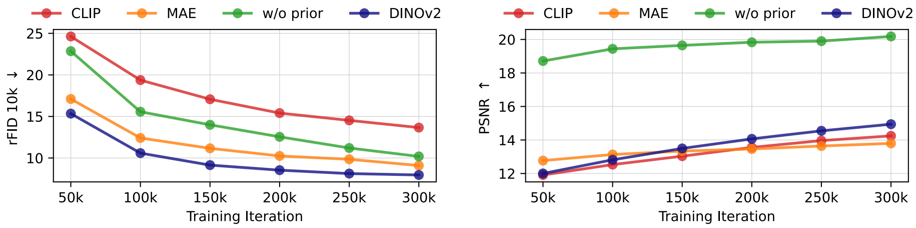


Figure S6: **Comparing SSL priors over training steps.** Our approach generalizes to different self-supervised methods. While the unregularized model without prior knowledge shows remarkable pixel-wise reconstruction, the latent space is not suitable for generation (see Table 4).

C Limitations

While our single-token representation enables highly efficient generation and significant compute savings, it may limit expressiveness in capturing fine-grained details, particularly for complex or high-resolution scenes. Extending our approach to support richer multi-token representations while preserving efficiency is an interesting direction for future work. While our experiments demonstrate that the single-token embedding preserves certain low-level spatial structures, achieving fine-grained control over object location and scene composition remains an open challenge.

Reconstruction-Generation trade-off Another limitation of our method lies in the trade-off imposed by cosine similarity regularization. While stronger regularization enhances the smoothness and structure of the latent space, which is crucial for stable generative modeling, it can also suppress low-level detail, leading to degraded pixel-wise reconstructions. This trade-off may limit the applicability of our approach in scenarios where very high visual reconstruction fidelity is critical.

Unleashing T2I for ImageNet-Pretrained Autoencoder We investigate the capabilities of our Image-trained encoder-decoder framework. Figure S10 shows qualitative text-to-image samples. Despite being trained exclusively on ImageNet, the latent space does not overfit and shows strong generalization, generating diverse and high-quality images that extend well beyond the ImageNet manifold. Although the model generates plausible images, we find it struggles with compositional prompts that require placing multiple objects within a scene (e.g., a cat and a dog side-by-side). This limitation is expected, since the object-centric bias of ImageNet offers little exposure to multi-object scenes. However, finetuning our encoder on more diverse data alleviates this issue and enables the generation of multi-object content.

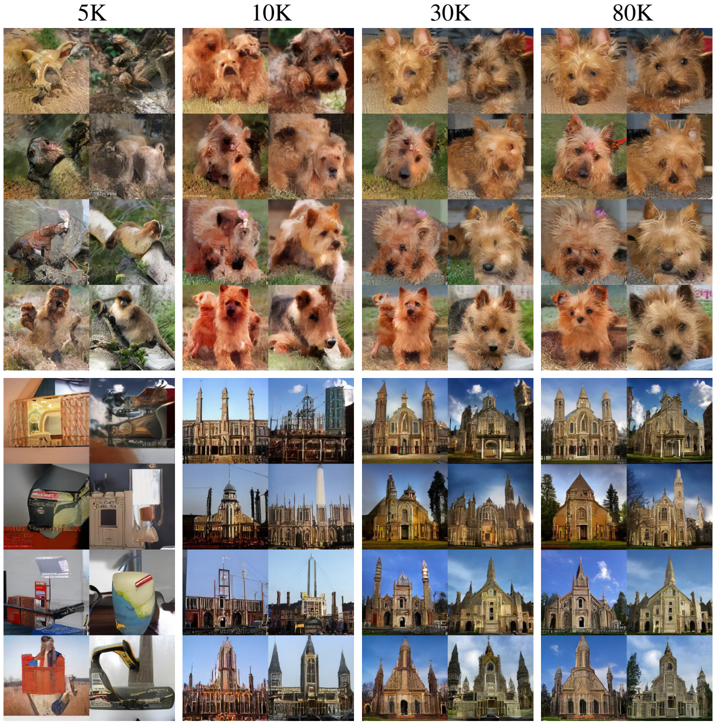


Figure S7: **Uncurated** class-conditional ImageNet generation results over training iterations (5k, 10k, 30k, and 80k). Note that our model produces good results as early as 30k training steps.

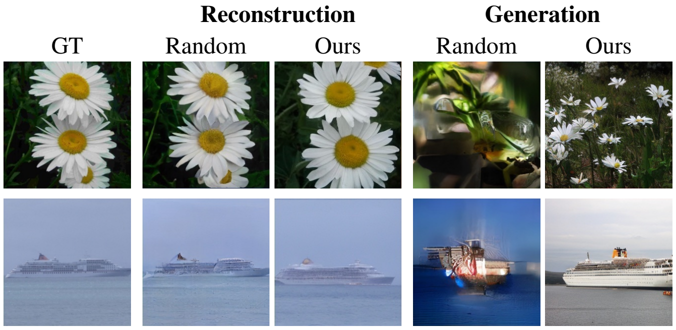
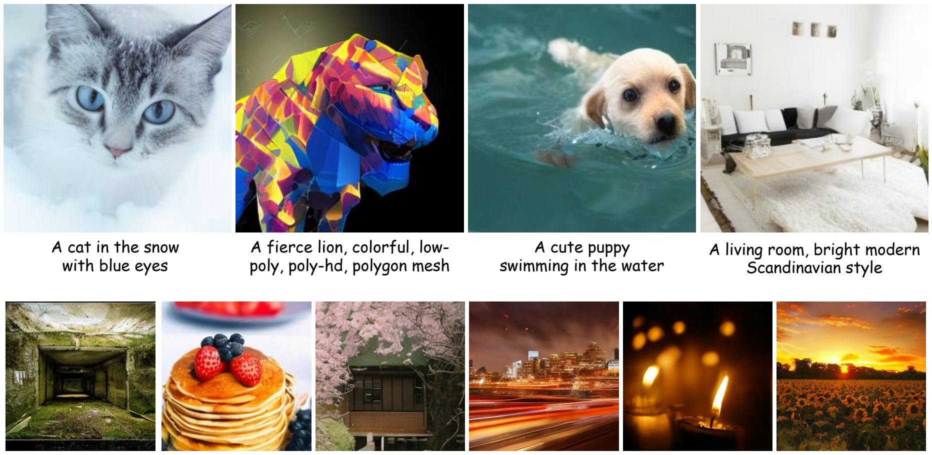


Figure S8: Qualitative comparison between a randomly-initialized encoder and ours. Generation refers to class-conditional samples with the same class as the corresponding GT image. While random initialization achieves stronger pixel-level reconstruction, it lacks the structured priors of pre-trained self-supervised encoders, resulting in poor generative performance. In contrast, our method balances reconstruction and generation.



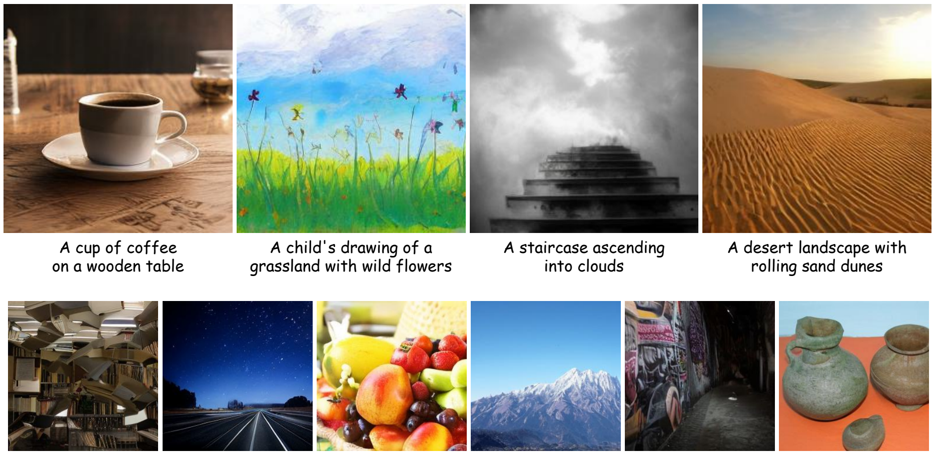
A cat in the snow with blue eyes

A fierce lion, colorful, low-poly, poly-hd, polygon mesh

A cute puppy swimming in the water

A living room, bright modern Scandinavian style

Figure S9: Additional text-to-image generation results with a CFG scale of 7.5 and RepTok encoder-decoder trained on the COYO dataset.



A cup of coffee on a wooden table

A child's drawing of a grassland with wild flowers

A staircase ascending into clouds

A desert landscape with rolling sand dunes

Figure S10: T2I generation results (CFG scale 3.5), using RepTok solely trained on ImageNet data with a latent space transformer. The autoencoder also transfers effectively to T2I tasks, producing visually compelling results.

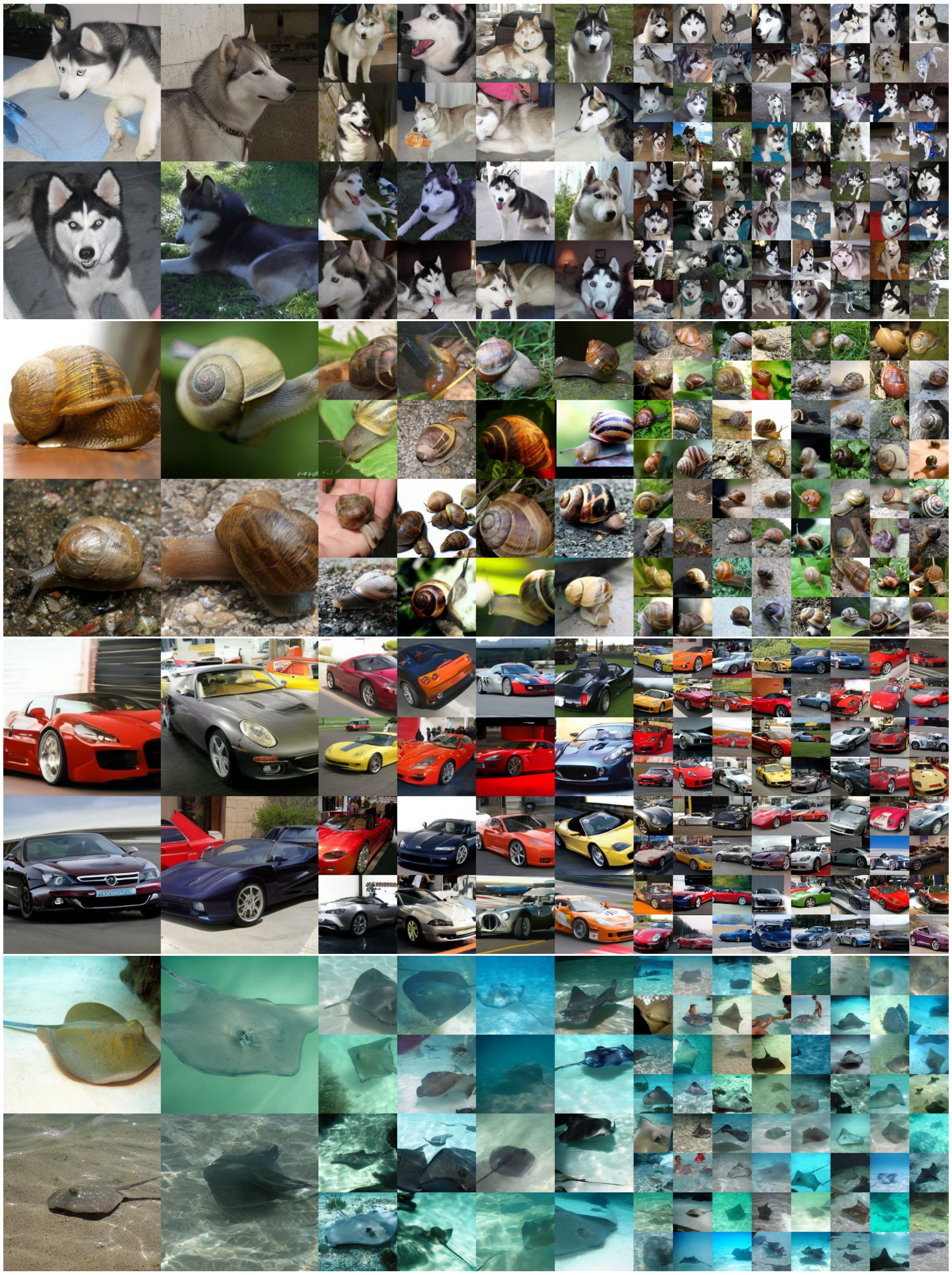


Figure S11: **Uncurated** class-conditional generation results of RepTok with CFG scale of 3.5.