CaTok: Taming Mean Flows for One-Dimensional Causal Image Tokenization

Supplementary Material

7. More implementation details

Table 5. Detailed configuration of CaTok-B and CaTok-L for tokenization and AR modeling.

Training Config	CaTok-B	CaTok-L	AR modeling
Optimizer	AdamW		
Peak learning rate	1×10^{-4}		5×10^{-5}
Minimum learning rate	0		
Learning rate schedule	cosine decay		constant
Batch size	1024		2048
Weight decay	0.05		
Epochs	80	160	400
Warmup epochs	0		96
Gradient clipping	3.0		
EMA	0.999		

Training setup follows [63], with detailed hyperparameters in Tab. 5. For reconstruction, we disable CFG in one-step sampling, and apply CFG with a scale of 2.0 in 25-step sampling. For 80-epoch training, we introduced the MeanFlow objective at epoch 10 and the selecting mechanism at epoch 40; for 160-epoch training, these corresponded to epochs 20 and 80, respectively. For generation, we do not use CFG with CaTok, and the CFG of AR model is the same as MUSE [5], MAR [29] and Semanticist [63], whici tunes down the guidance scale of small-indexed tokens to improve the diversity of generated sample.

8. More experiments

Table 6. **Reconstruction and generation results of CaTok-VQ**

Method	#Param.	Token	Step	rFID	gFID
LlamaGen	343M	256	256	2.19	3.80
CaTok-LlamaGen	343M	256	256	3.81	3.35

CaTok-VQ. To evaluate the effectiveness of our approach with VQ and to avoid the cumulative errors introduced by causal MAR [29] when the number of tokens increases, we conduct a straightforward comparison experiment with LlamaGen [51]. Specifically, we integrate FSQ into CaTok without modifying any part of the original training recipe. As shown in Tab. 6, because we do not perform hyperparameter tuning tailored for VQ training, nor incorporate additional techniques such as perceptual losses or post-training [47, 51], CaTok-VQ performs significantly worse than LlamaGen's tokenizer in terms of rFID (3.81 *vs.* 2.19). However, due to the inherent causality of CaTok-VQ's visual tokens—which is advantageous for autoregressive modeling—its AR generation performance surpasses that of LlamaGen (3.35 *vs.* 3.80), which further demonstrates the superior effectiveness of our approach. We believe that improved training of the VQ tokenizer, along with larger autoregressive models, can lead to further gains in generation performance.

(w/o CFG)	Learned tokens in tokenization training	Used tokens in AR modeling	gFID	IS
Semanticist-L-256	256	256	7.60	121.5
CaTok-L-256	256	256	**5.52**	**153.9**
Semanticist-L-256	256	32	4.96	147.4
CaTok-L-256†	256	32	**4.77**	**165.2**

Apple-to-apple comparison with Semanticist without CFG. We conduct a comparison with Semanticist under matched settings without CFG: we train an AR model using the official Semanticist tokenizer checkpoint with 256 tokens and directly evaluate the official 32-token checkpoint in the no-CFG setting. †: we freeze the ViT encoder and fine-tune the DiT with nested dropout.

Extensions on the REPA encoder, latent space, and image resolution. CaTok exhibits consistent reconstruction behavior across different REPA teacher [48, 53] and latent spaces [15, 29], and the reconstruction drop mainly stems from DiT re-compressing the latent space and can be alleviated by reducing the DiT patch size to smaller. Moreover, the DiT can naturally generalize to higher resolutions via a training-free patchwise diffuse-and-blend strategy (as in FlowMo [47]). Since MAR-VAE is not trained at 512×512, we replace it with SD-VAE for training at 512 resolution, and adopt the high-resolution timestep shift used in SD3.

CaTok-B-256 on ImageNet	rFID	PSNR	SSIM
DINOv3	1.16	22.43	0.674
SigLIP2	1.12	21.96	0.657
MAR-VAE w/ DiT-B/2	0.99	22.69	0.672
SD-VAE	1.34	21.99	0.658
512 resolution (training-free)	0.60	27.74	0.778
512 resolution (trained w/ SD-VAE)	1.07	24.92	0.705

Additional dataset. To further evaluate the generalization ability of CaTok beyond ImageNet, we conduct additional experiments on the COCO-val-5K dataset [30]. CaTok achieves consistent performance on COCO-val-5K, indicating that the learned representations generalize well to datasets with different distributions.

Models on COCO-5K	rFID	PSNR	SSIM
LlamaGen-16x16	8.11	20.42	0.678
Semanticist-L-256	5.64	21.36	0.640
CATOK-L-256	**4.78**	**22.43**	**0.690**

Non-autoregressive generator. To further study the applicability of CATOK under different generation paradigms, we replace the autoregressive generator with a non-autoregressive generator, ϵMaskGiT, and evaluate the model under 8-step sampling. Under this setting, CATOK achieves better performance than TiTok, indicating that the proposed tokenizer is compatible with both autoregressive and mask-based generation. Furthermore, a variant of CATOK without token dropout still yields improved performance, suggesting that the introduced visual causality also benefits non-autoregressive modeling.

Tokenizer	Generator	Step	gFID	IS
TiTok-L-32	MaskGiT-L	8	2.77	194.0
CATOK-L-32 w/o causality	ϵMaskGiT-L	8	3.26	210.4
CATOK-L-32	ϵMaskGiT-L	8	**2.69**	**223.7**