Language-Guided Image Tokenization for Generation


Kaiwen Zha


Lijun Yu


Alireza Fathi


David Ross


Cordelia Schmid


Dina Katabi


Xiuye Gu

DeepMind

MIT CSAIL

CVPR 2025

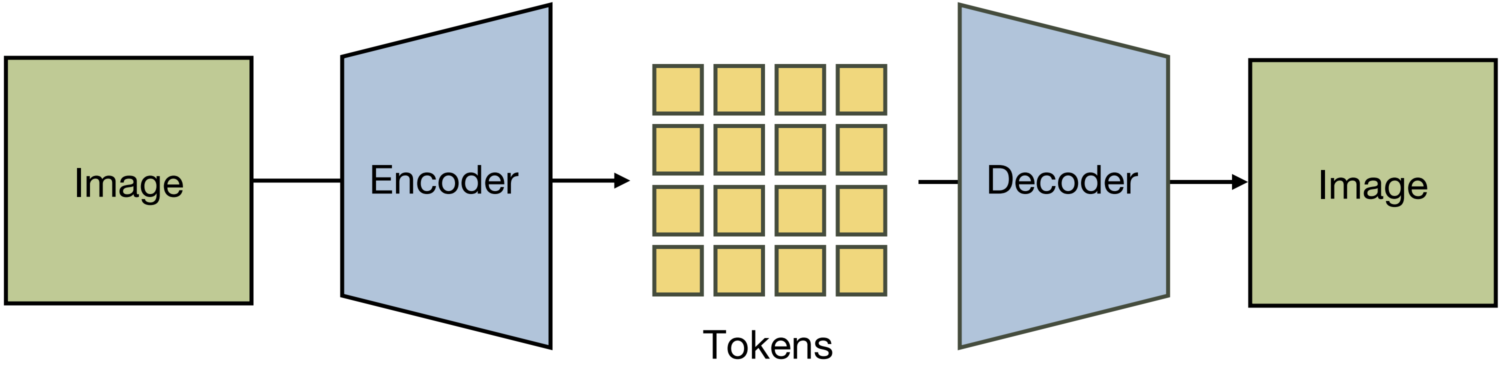
https://kaiwenzha.github.io/textok/

Image Generation Made Great Progress

*Examples generated by FLUX (https://www.reddit.com/r/StableDiffusion/comments/1ehiz51/flux_image_examples/)

Tokenization is Key to Image Generation

Tokenization: Compresses raw image data into a compact low-dimensional latent representation (we call it “token”) through training an autoencoder



Problem: Tradeoff between Compression and Quality

- **High** compression rate:

  Low computational cost, bad reconstruction quality

- **Low** compression rate:

  Good reconstruction quality, high computational cost

Can we achieve the best of both worlds,
i.e., **low cost** and **high quality**?

Our idea: Use Text during Tokenization

Tokenization: Finding a compact and comprehensive representation of an image

The most compact and comprehensive representation available of an image is its **caption**.

Our idea: Use Text during Tokenization

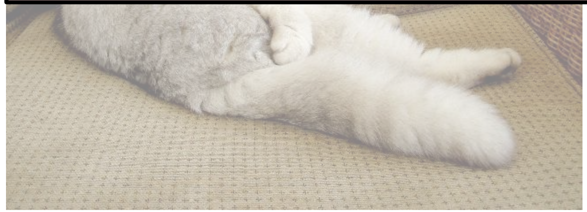
Tokenization: Finding a compact and comprehensive representation of an image

A fluffy, silver-and-white Persian cat lounges comfortably on a beige, textured sofa, its long, luxurious fur creating a soft cloud-like texture. The cat's round body and plush tail are relaxed, its paws tucked gently beneath it, and its green eyes are partially visible in a somewhat pensive expression. It appears content and at ease in its domestic environment.

Our idea: Use Text during Tokenization

Tokenization: Finding a compact and comprehensive representation of an image

Using text (i.e., image caption) during tokenization can **simplify semantic learning.**



s
ng,
ure.
ed, its
paws tucked gently beneath it, and its green eyes are partially visible in a somewhat pensive expression. It appears content and at ease in its domestic environment.

Our idea: Use Text during Tokenization

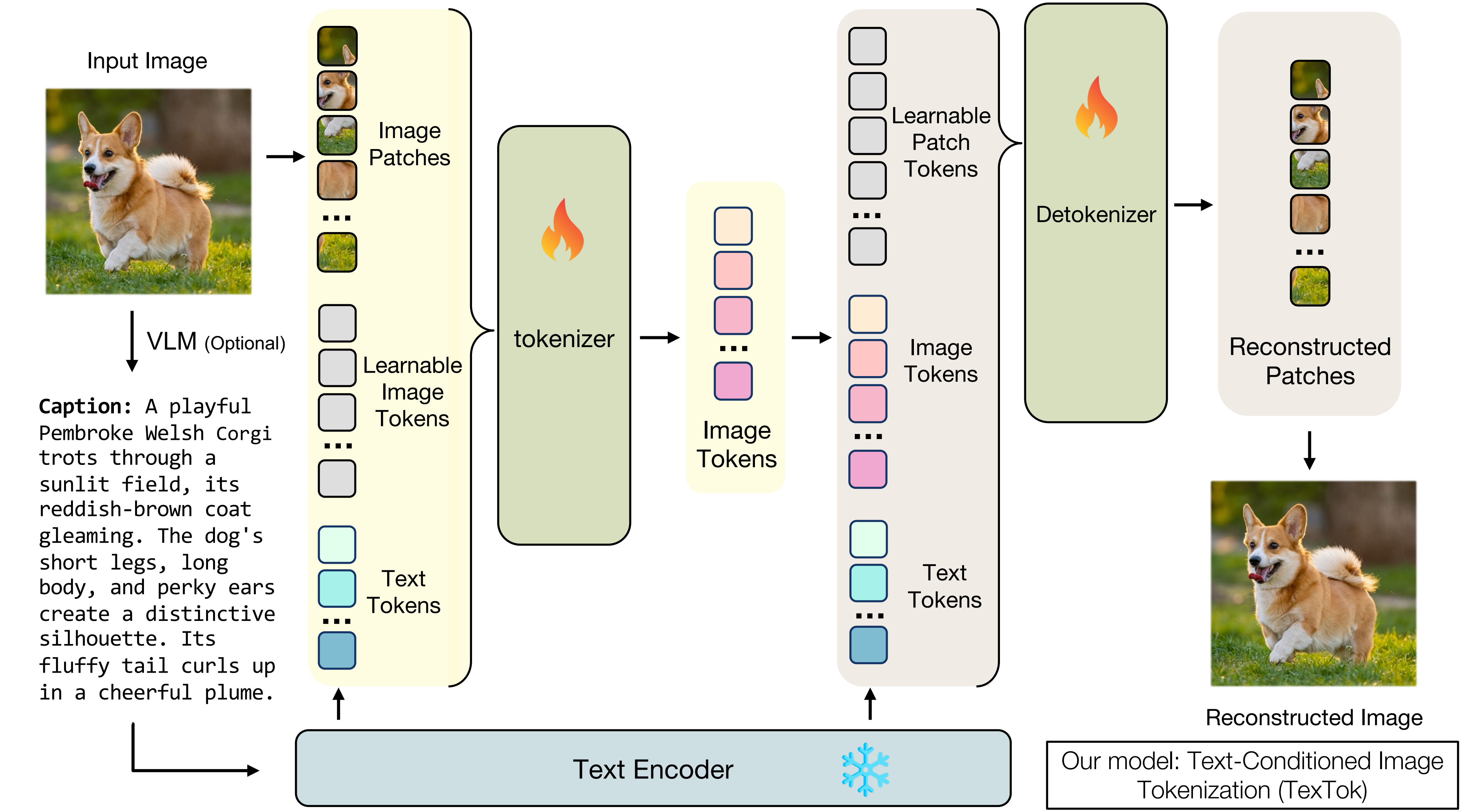
Tokenization: Finding a compact and comprehensive representation of an image

paws tucked gently beneath it, and its green eyes
are partially visible in a somewhat pensive

Using text (i.e., image caption) during tokenization can **simplify semantic learning.**

Achieve better quality without compromising cost!

Our model: Text-Conditioned Image Tokenization (TexTok)



Input Image
VLM (Optional)
Caption: A playful Pembroke Welsh Corgi trots through a sunlit field, its reddish-brown coat gleaming. The dog's short legs, long body, and perky ears create a distinctive silhouette. Its fluffy tail curls up in a cheerful plume.
Image Patches
Learnable Image Tokens
Text Tokens
tokenizer
Image Tokens
Learnable Patch Tokens
Image Tokens
Text Tokens
Detokenizer
Reconstructed Patches
Reconstructed Image
Text Encoder
Our model: Text-Conditioned Image Tokenization (TexTok)

Reconstruction Results

Input Image

Reconstruction Results

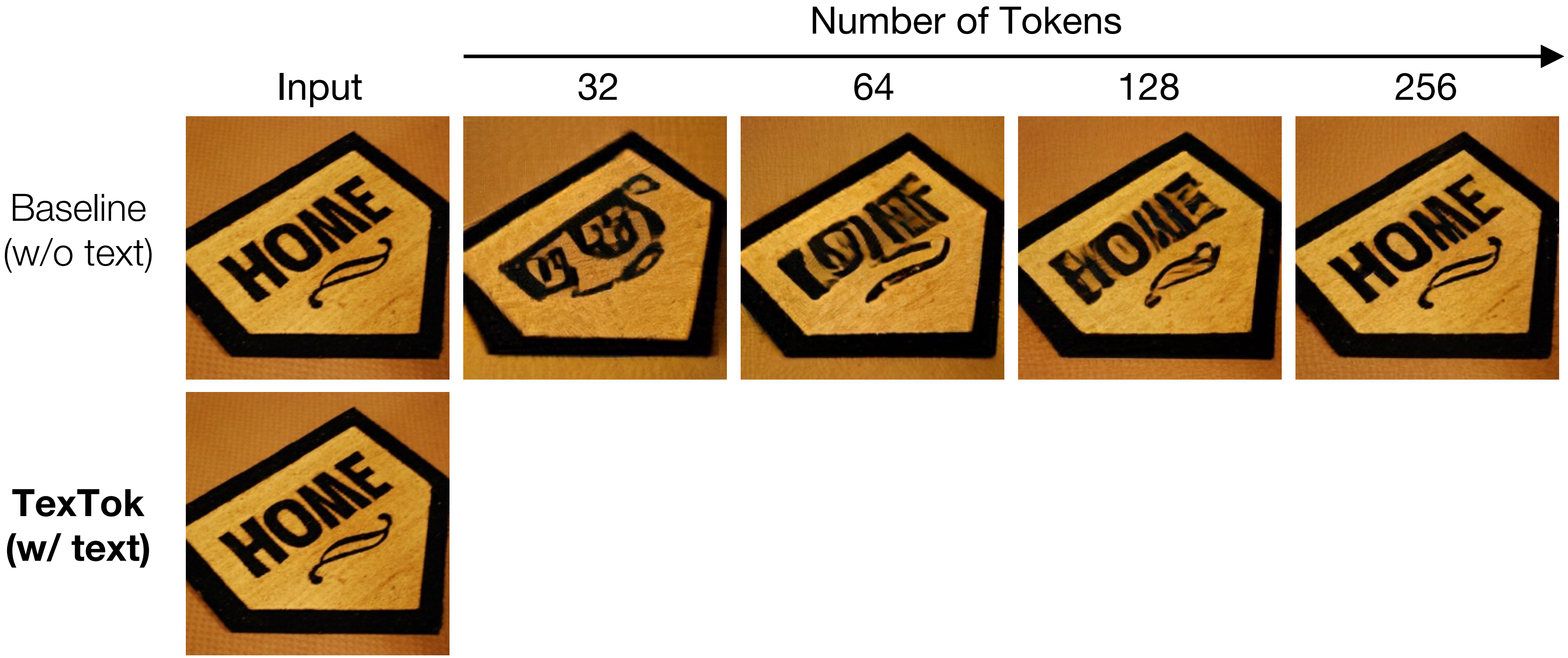
Number of Tokens

Input | 32 | 64 | 128 | 256

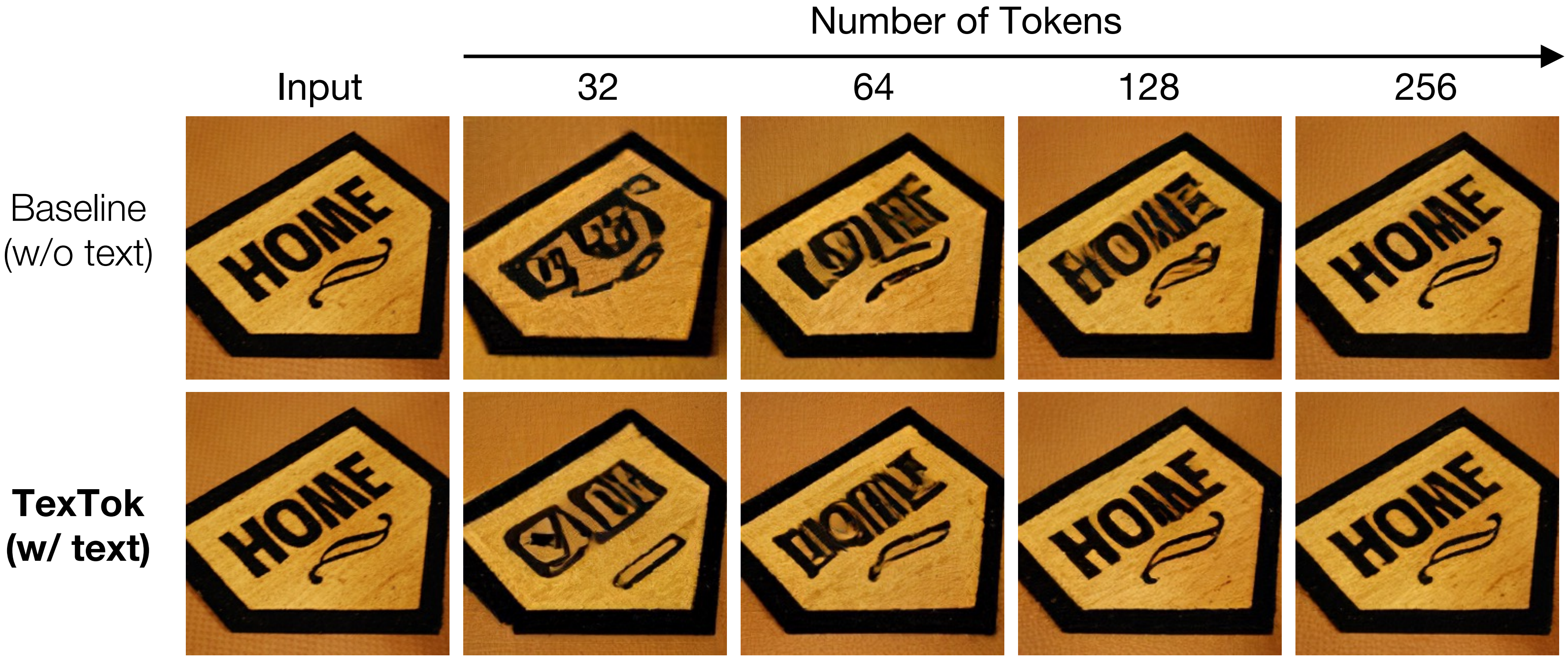
Baseline (w/o text)

TexTok (w/ text)

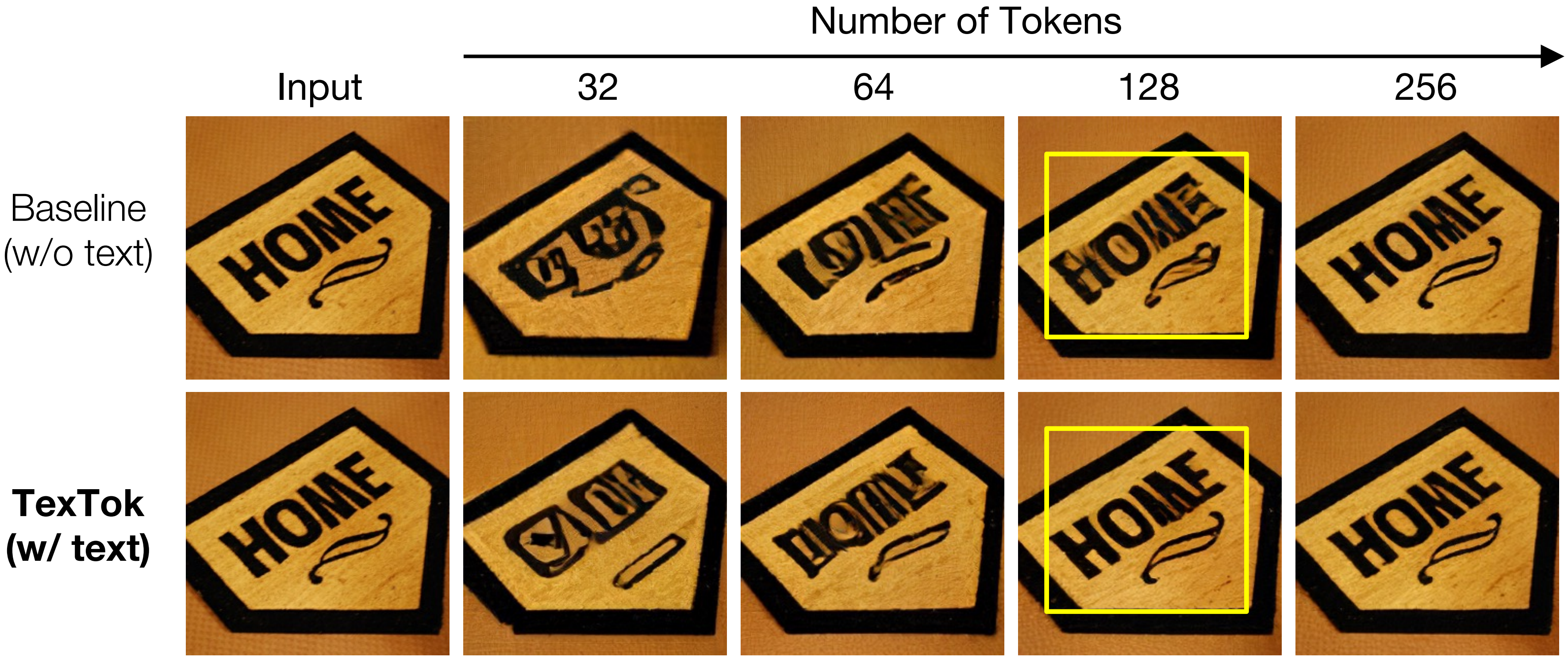
Reconstruction Results



Reconstruction Results



Reconstruction Results

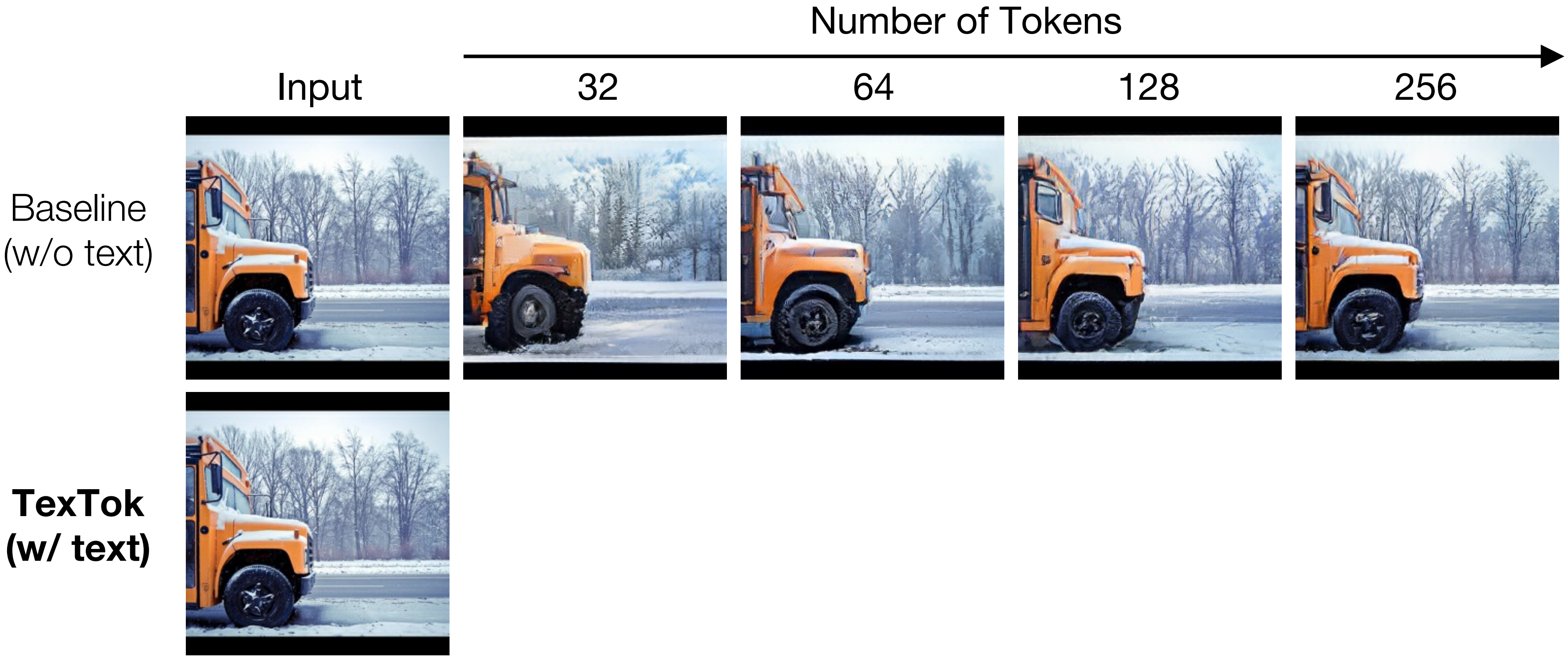


Reconstruction Results

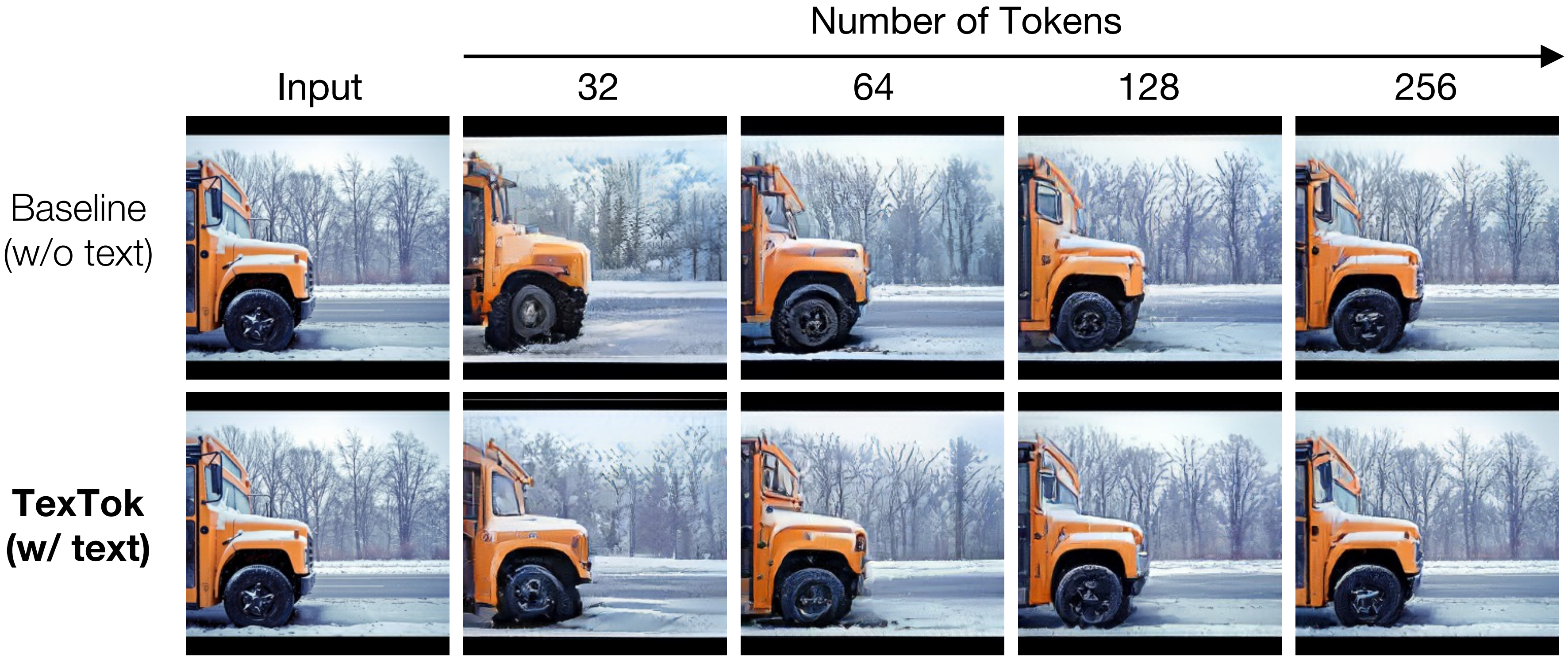
Reconstruction Results

Input Image

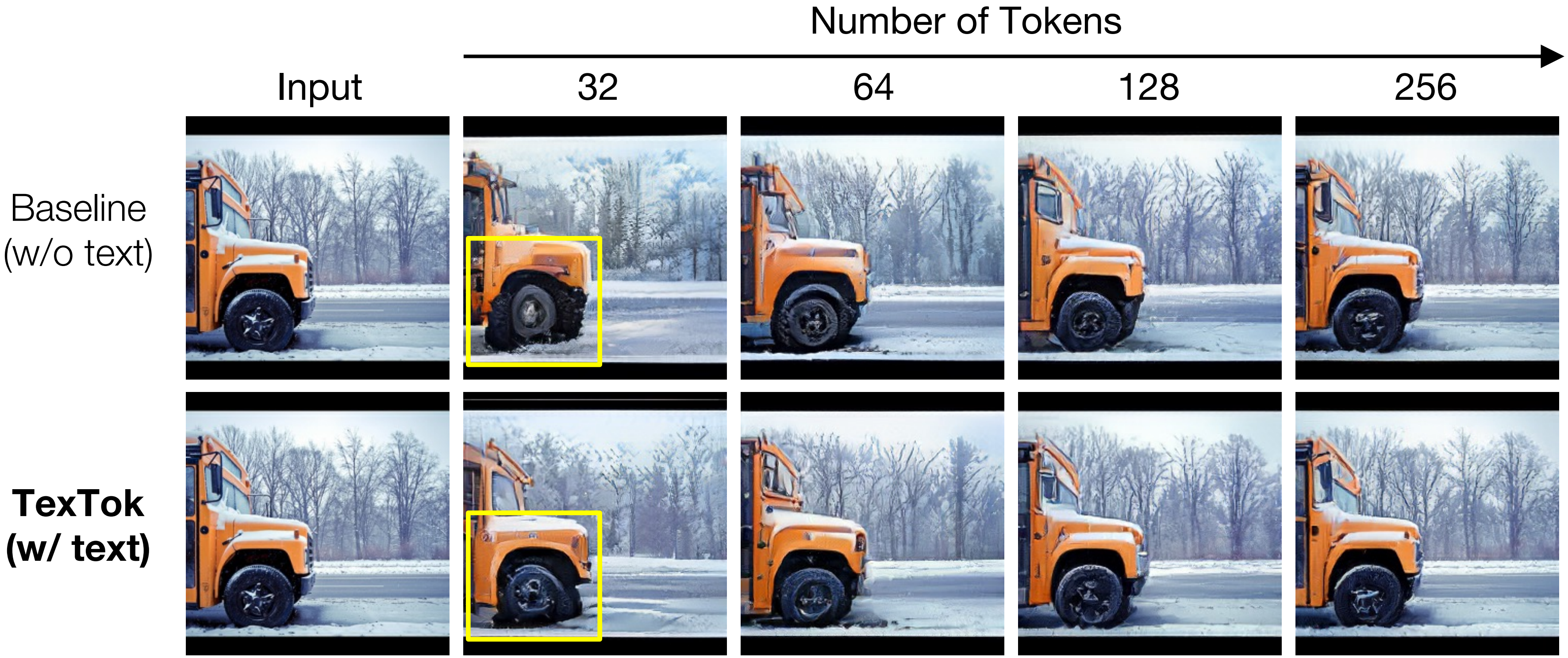
Reconstruction Results



Reconstruction Results

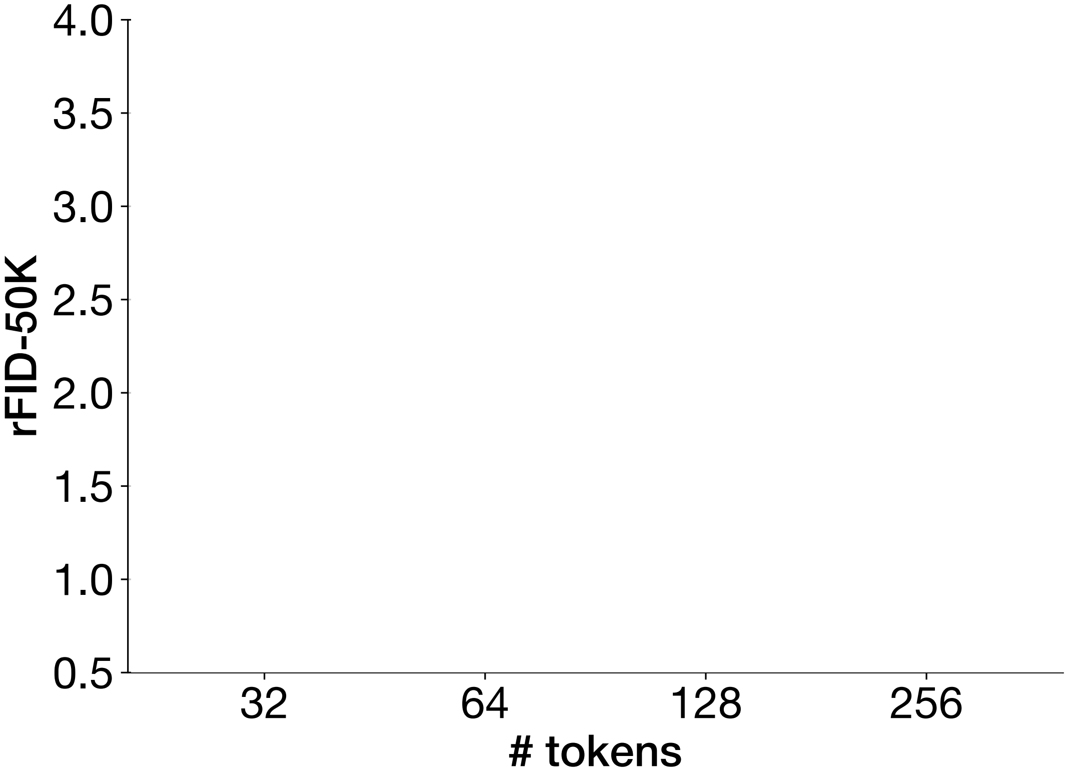


Reconstruction Results



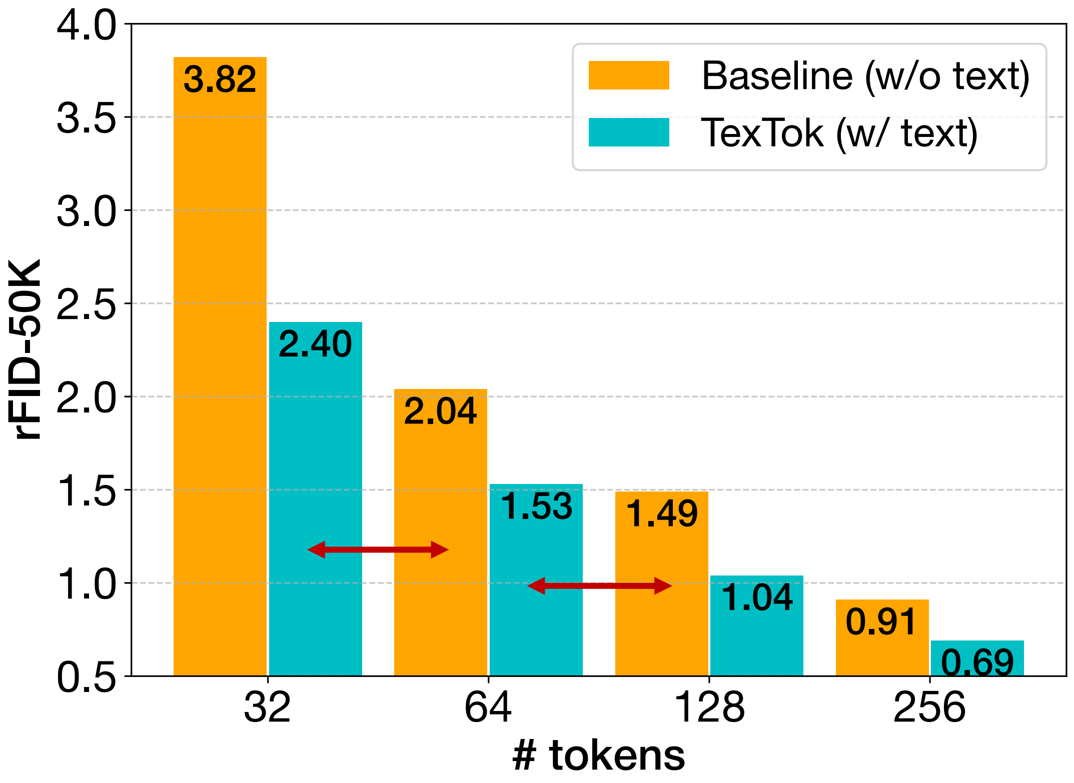
Quantitative Reconstruction Results

ImageNet 256x256



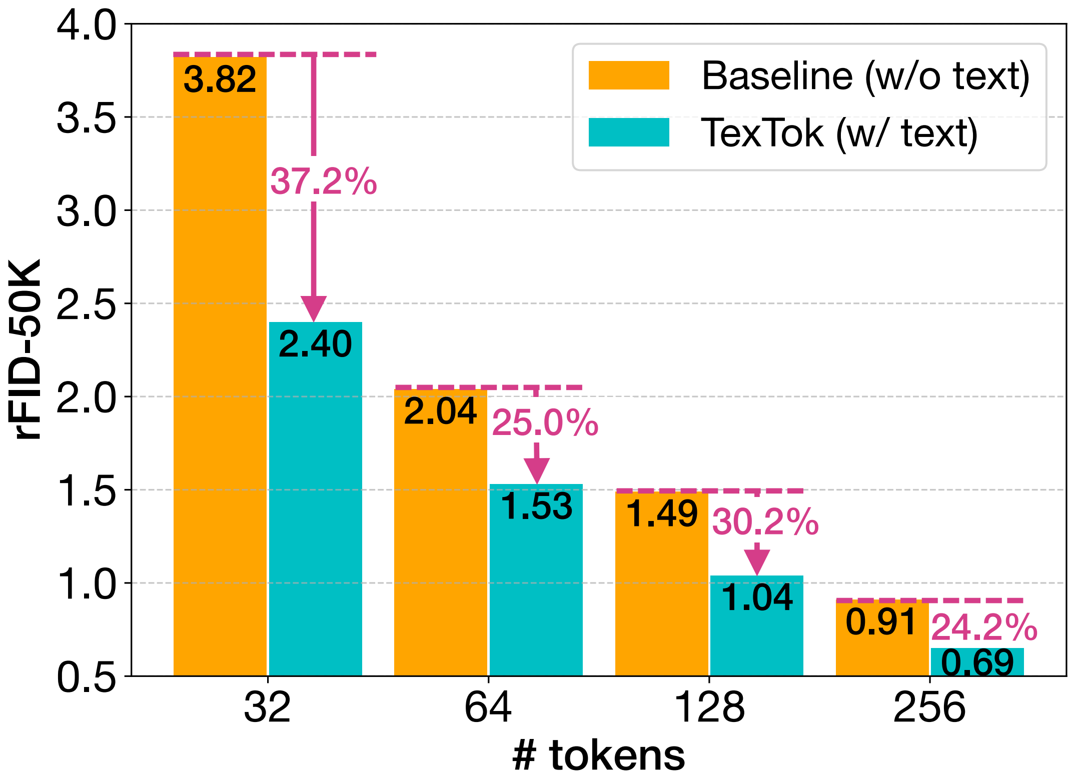
Better Quality at High Compression

ImageNet 256x256



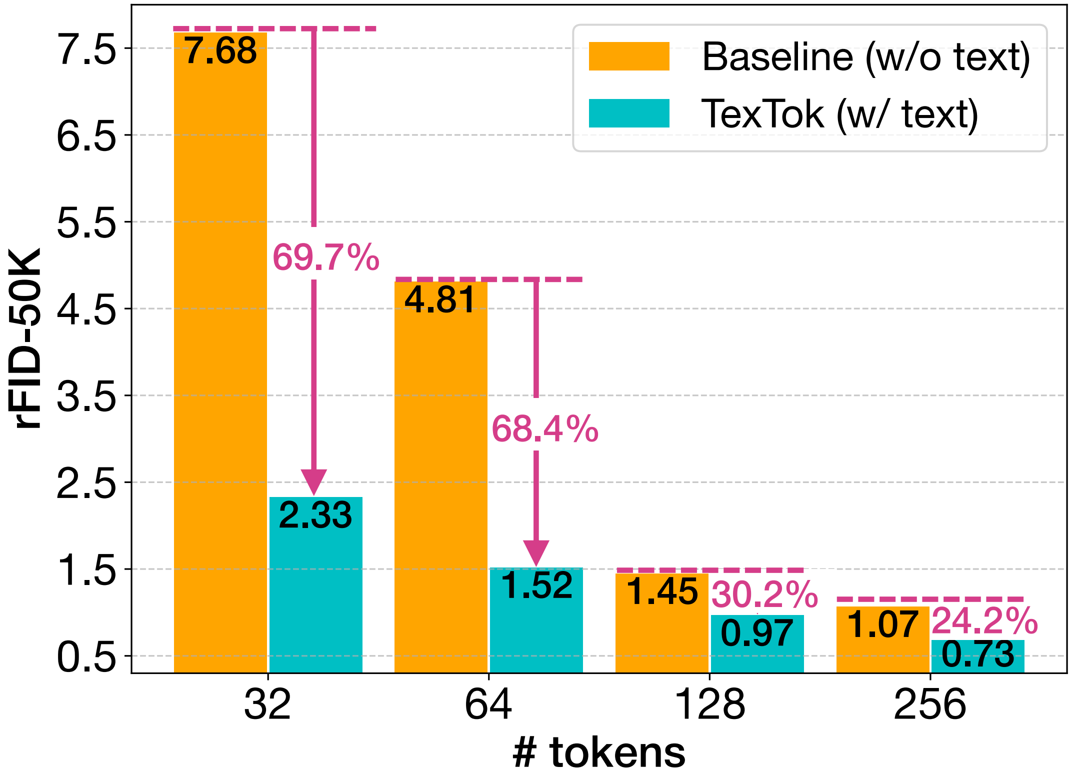
Better Quality at High Compression

ImageNet 256x256



Better Quality at High Compression

ImageNet 512x512



More Detailed Reconstruction Results

Reconstruction						
tokenizer	# tokens	rFID ↓	rIS↑	PSNR↑	SSIM ↑	LPIPS↓
(a) ImageNet 256×256						
SD-VAE-f8 [37]	1024 (d=4)	1.20†	-	-	-	-
Baseline-32 (w/o text)	32 (d=8)	3.82	117.1	17.67	0.4281	0.3270
TexTok-32 (w/ text)		**2.40**	**156.2**	**18.32**	**0.4463**	**0.2884**
Baseline-64 (w/o text)	64 (d=8)	2.04	147.2	19.52	0.4801	0.2343
TexTok-64 (w/ text)		**1.53**	**169.8**	**20.10**	**0.4971**	**0.2126**
Baseline-128 (w/o text)	128 (d=8)	1.49	160.5	20.51	0.5102	0.1913
TexTok-128 (w/ text)		**1.04**	**183.3**	**22.05**	**0.5618**	**0.1499**
Baseline-256 (w/o text)	256 (d=8)	0.91	178.3	23.05	0.5950	0.1225
TexTok-256 (w/ text)		**0.69**	**192.6**	**24.38**	**0.6454**	**0.0998**

Tokenization Improvements Translate to Generation

Reconstruction							Generation	
tokenizer	# tokens	rFID ↓	rIS↑	PSNR↑	SSIM ↑	LPIPS↓	gFID ↓	gIS ↑
(a) ImageNet 256×256								
SD-VAE-f8 [37]	1024 (d=4)	1.20†	-	-	-	-	9.62	121.5
Baseline-32 (w/o text)	32 (d=8)	3.82	117.1	17.67	0.4281	0.3270	4.97	170.3
TexTok-32 (w/ text)		**2.40**	**156.2**	**18.32**	**0.4463**	**0.2884**	**3.55**	**205.3**
Baseline-64 (w/o text)	64 (d=8)	2.04	147.2	19.52	0.4801	0.2343	3.30	188.9
TexTok-64 (w/ text)		**1.53**	**169.8**	**20.10**	**0.4971**	**0.2126**	**2.88**	**209.2**
Baseline-128 (w/o text)	128 (d=8)	1.49	160.5	20.51	0.5102	0.1913	3.19	190.1
TexTok-128 (w/ text)		**1.04**	**183.3**	**22.05**	**0.5618**	**0.1499**	**2.75**	**210.9**
Baseline-256 (w/o text)	256 (d=8)	0.91	178.3	23.05	0.5950	0.1225	2.91	197.2
TexTok-256 (w/ text)		**0.69**	**192.6**	**24.38**	**0.6454**	**0.0998**	**2.68**	**219.6**

System-level Generation Benchmarking

			(a) ImageNet 256×256					(b) ImageNet 512×512				
Model	#Params (G)	#Params (T)	FID↓	IS↑	Precision↑	Recall↑	#tokens	FID↓	IS↑	Precision↑	Recall↑	#tokens
latent diffusion												
LDM-4 [37]	400M	55M	3.60	247.7	0.87	0.48	4096 (d=3)	-	-	-	-	-
U-ViT-H [2]	501M	84M	2.29	263.9	0.82	0.57	1024* (d=4)	4.05	263.8	0.84	0.48	4096* (d=4)
DiT-XL/2 [32]	675M	84M	2.27	278.2	0.83	0.57	1024* (d=4)	3.04	240.8	0.84	0.54	4096* (d=4)
DiffiT [14]	-	-	1.73	276.5	0.80	0.62	-	2.67	252.1	0.83	0.55	-
MDTv2-XL/2 [12]	676M	84M	1.58	314.7	0.79	0.65	1024* (d=4)	-	-	-	-	-
REPA + SiT-XL/2 [51]	675M	84M	1.80	284.0	0.81	0.61	1024* (d=4)	-	-	-	-	-
EDM2-XXL [21]	1.5B	84M	-	-	-	-	-	1.81	-	-	-	4096 (d=4)
Ours												
TexTok-32 + DiT-XL	675M	176M	2.75	294.6	0.83	0.56	32 (d=8)	2.74	303.2	0.83	0.56	32 (d=8)
TexTok-64 + DiT-XL	675M	176M	2.06	290.0	0.81	0.60	64 (d=8)	1.99	301.9	0.82	0.6	64 (d=8)
TexTok-128 + DiT-XL	675M	176M	1.66	294.4	0.80	0.61	128 (d=8)	1.80	305.4	0.81	0.63	128 (d=8)
TexTok-256 + DiT-XL	675M	176M	**1.46**	303.1	0.79	0.64	256 (d=8)	**1.62**	313.8	0.80	0.64	256 (d=8)

Our model achieves state-of-the-art generation performance at the time of submission

System-level Generation Benchmarking

Model	#Params (G)	#Params (T)	(a) ImageNet 256×256					(b) ImageNet 512×512				
			FID↓	IS↑	Precision↑	Recall↑	#tokens	FID↓	IS↑	Precision↑	Recall↑	#tokens
latent diffusion												
LDM-4 [37]	400M	55M	3.60	247.7	0.87	0.48	4096 (d=3)	-	-	-	-	-
U-ViT-H [2]	501M	84M	2.29	263.9	0.82	0.57	1024* (d=4)	4.05	263.8	0.84	0.48	4096* (d=4)
DiT-XL/2 [32]	675M	84M	2.27	278.2	0.83	0.57	1024* (d=4)	3.04	240.8	0.84	0.54	4096* (d=4)
DiffiT [14]	-	-	1.73	276.5	0.80	0.62	-	2.67	252.1	0.83	0.55	-
MDTv2-XL/2 [12]	676M	84M	1.58	314.7	0.79	0.65	1024* (d=4)	-	-	-	-	-
REPA + SiT-XL/2 [51]	675M	84M	1.80	284.0	0.81	0.61	1024* (d=4)	-	-	-	-	-
EDM2-XXL [21]	1.5B	84M	-	-	-	-	-	1.81	-	-	-	4096 (d=4)
Ours												
TexTok-32 + DiT-XL	675M	176M	2.75	294.6	0.83	0.56	32 (d=8)	2.74	303.2	0.83	0.56	32 (d=8)
TexTok-64 + DiT-XL	675M	176M	2.06	290.0	0.81	0.60	64 (d=8)	1.99	301.9	0.82	0.6	64 (d=8)
TexTok-128 + DiT-XL	675M	176M	1.66	294.4	0.80	0.61	128 (d=8)	1.80	305.4	0.81	0.63	128 (d=8)
TexTok-256 + DiT-XL	675M	176M	**1.46**	303.1	0.79	0.64	256 (d=8)	**1.62**	313.8	0.80	0.64	256 (d=8)

Our model with 64 tokens perform better than vanilla DiT with 1024 tokens on ImageNet-256

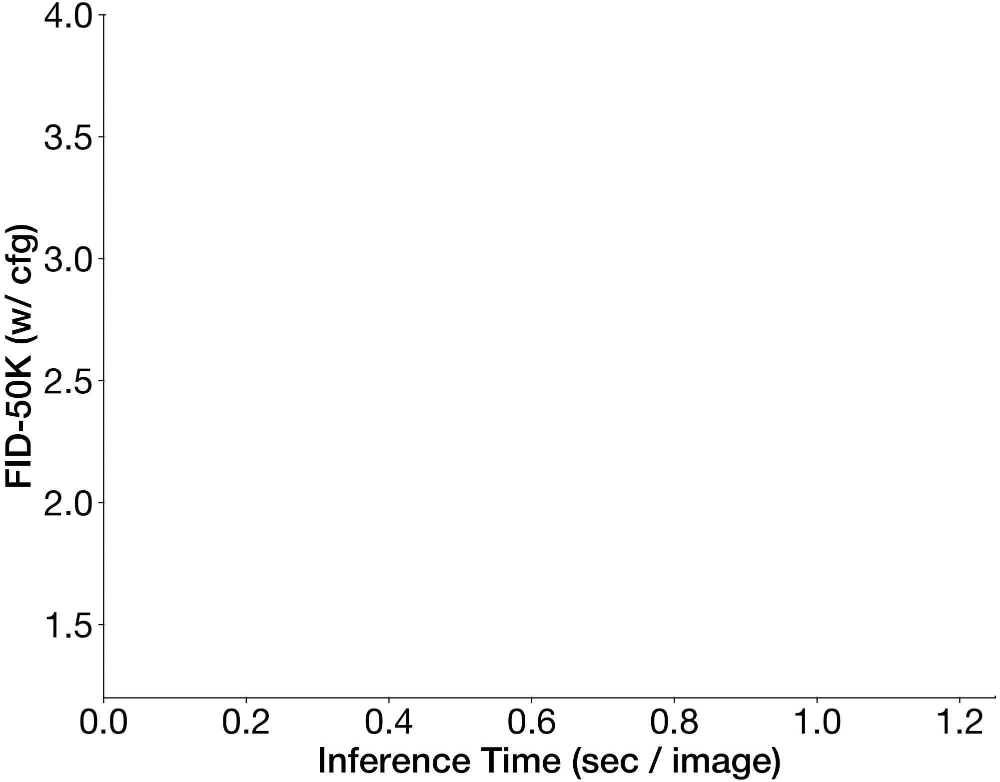
System-level Generation Benchmarking

			(a) ImageNet 256×256					(b) ImageNet 512×512				
Model	#Params (G)	#Params (T)	FID↓	IS↑	Precision↑	Recall↑	#tokens	FID↓	IS↑	Precision↑	Recall↑	#tokens
latent diffusion												
LDM-4 [37]	400M	55M	3.60	247.7	0.87	0.48	4096 (d=3)	-	-	-	-	-
U-ViT-H [2]	501M	84M	2.29	263.9	0.82	0.57	1024* (d=4)	4.05	263.8	0.84	0.48	4096* (d=4)
DiT-XL/2 [32]	675M	84M	2.27	278.2	0.83	0.57	1024* (d=4)	3.04	240.8	0.84	0.54	4096* (d=4)
DiffiT [14]	-	-	1.73	276.5	0.80	0.62	-	2.67	252.1	0.83	0.55	-
MDTv2-XL/2 [12]	676M	84M	1.58	314.7	0.79	0.65	1024* (d=4)	-	-	-	-	-
REPA + SiT-XL/2 [51]	675M	84M	1.80	284.0	0.81	0.61	1024* (d=4)	-	-	-	-	-
EDM2-XXL [21]	1.5B	84M	-	-	-	-	-	1.81	-	-	-	4096 (d=4)
Ours												
TexTok-32 + DiT-XL	675M	176M	2.75	294.6	0.83	0.56	32 (d=8)	2.74	303.2	0.83	0.56	32 (d=8)
TexTok-64 + DiT-XL	675M	176M	2.06	290.0	0.81	0.60	64 (d=8)	1.99	301.9	0.82	0.6	64 (d=8)
TexTok-128 + DiT-XL	675M	176M	1.66	294.4	0.80	0.61	128 (d=8)	1.80	305.4	0.81	0.63	128 (d=8)
TexTok-256 + DiT-XL	675M	176M	**1.46**	303.1	0.79	0.64	256 (d=8)	**1.62**	313.8	0.80	0.64	256 (d=8)

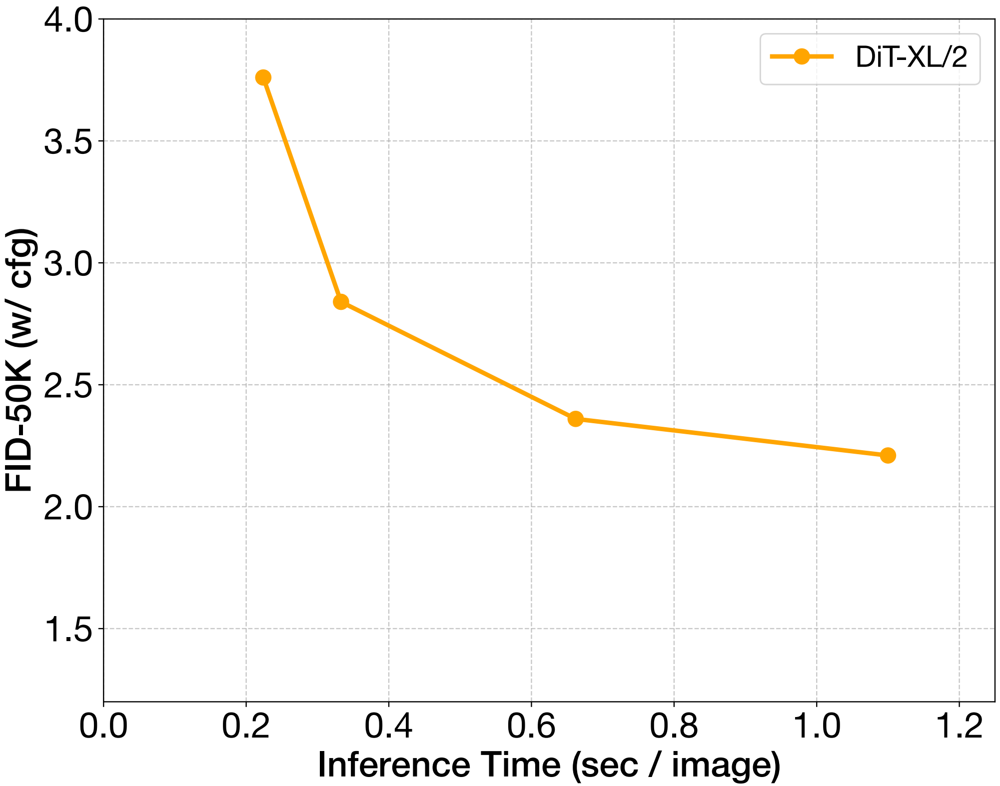
Our model with 32 tokens perform better than vanilla DiT with 4096 tokens on ImageNet-512

Improved Generation Speed/Quality Frontier

ImageNet 256x256

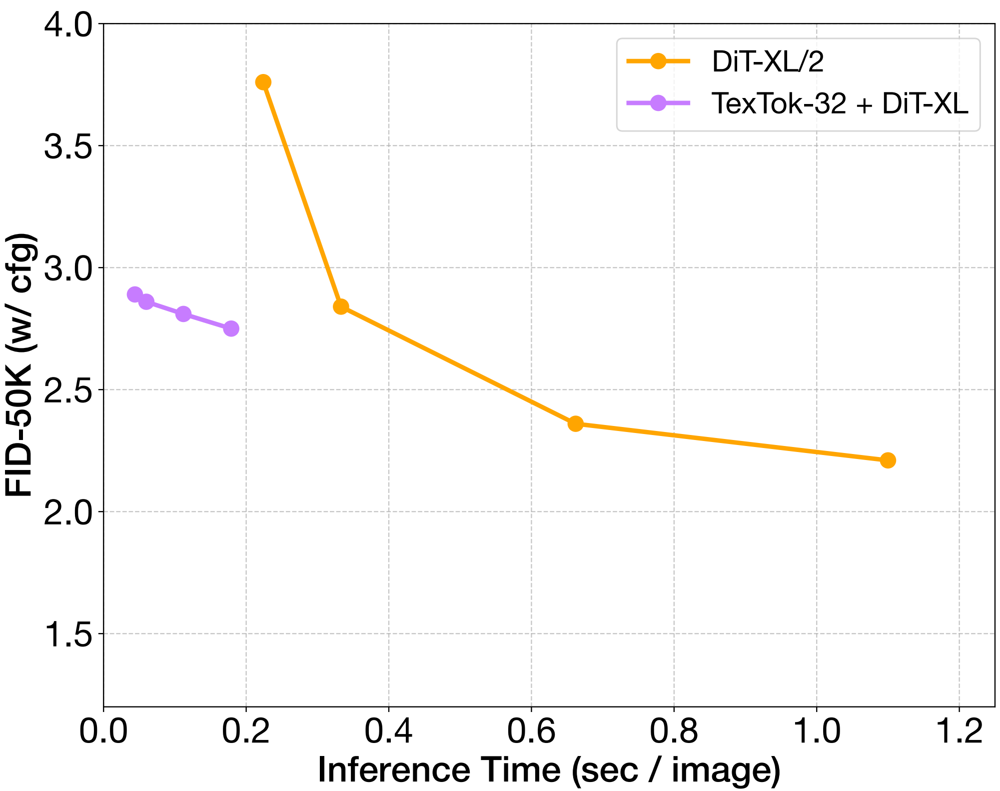


Improved Generation Speed/Quality Frontier



ImageNet 256x256

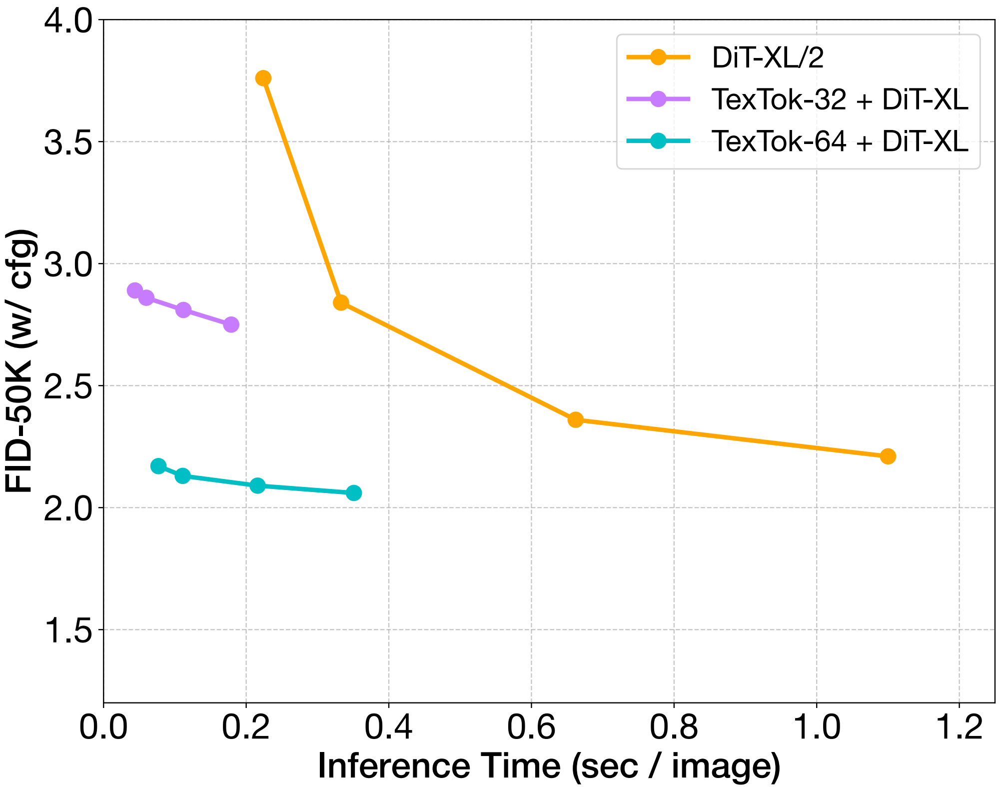
Improved Generation Speed/Quality Frontier



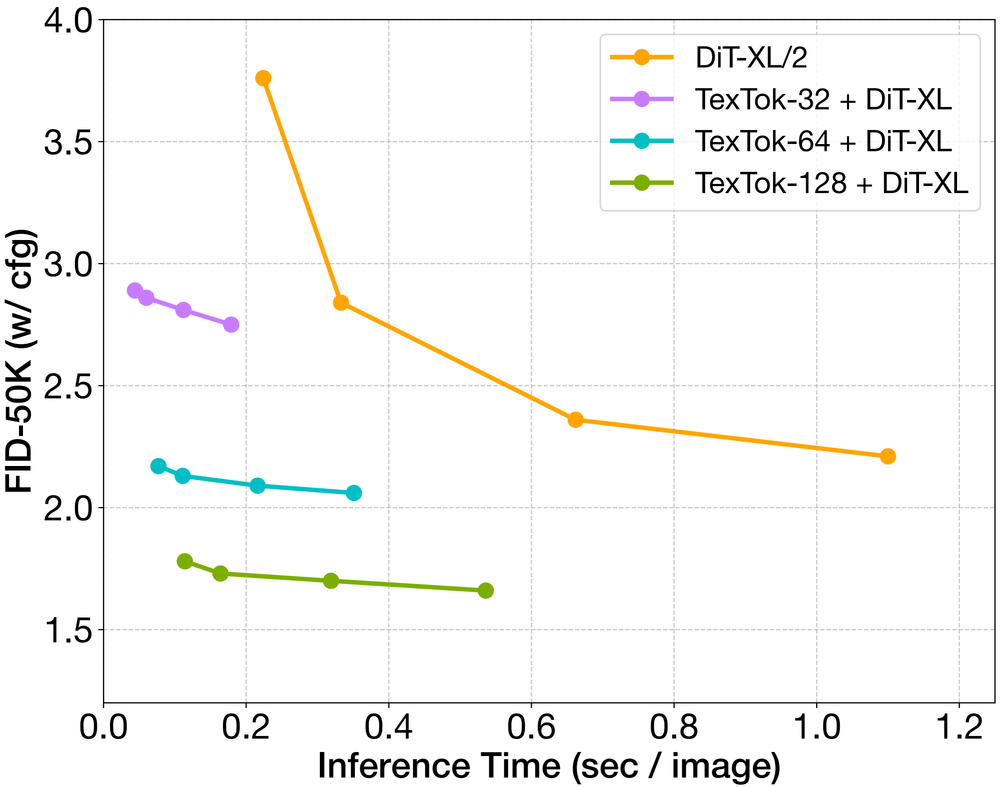
ImageNet 256x256

Improved Generation Speed/Quality Frontier

ImageNet 256x256



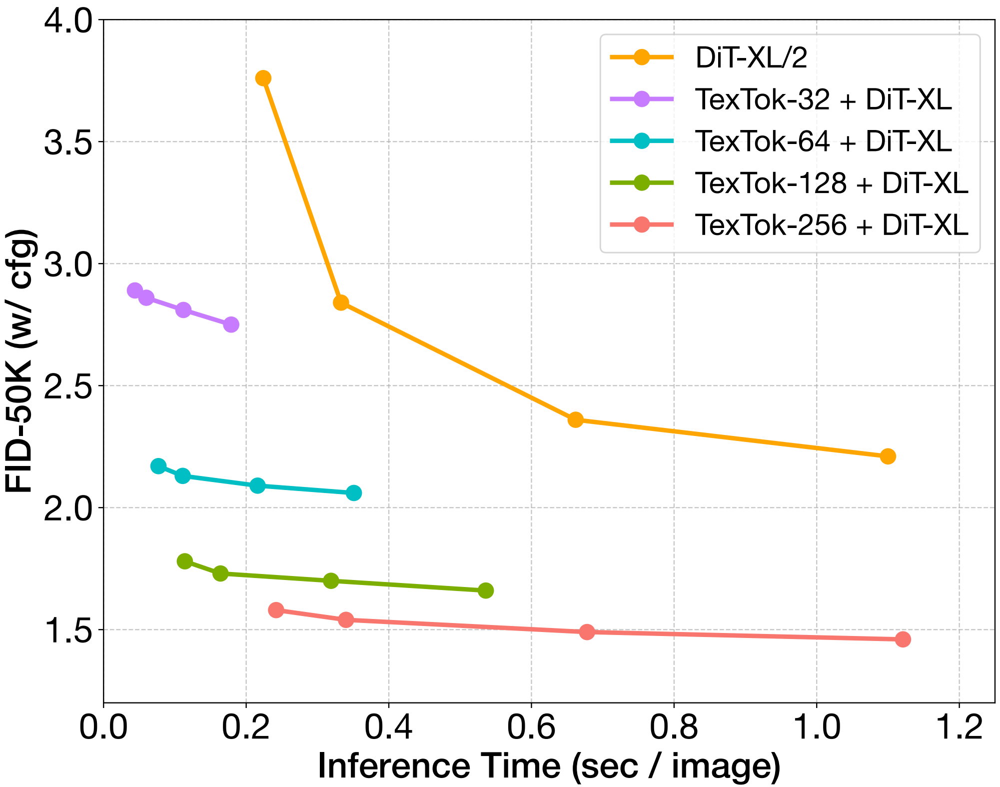
Improved Generation Speed/Quality Frontier



ImageNet 256x256

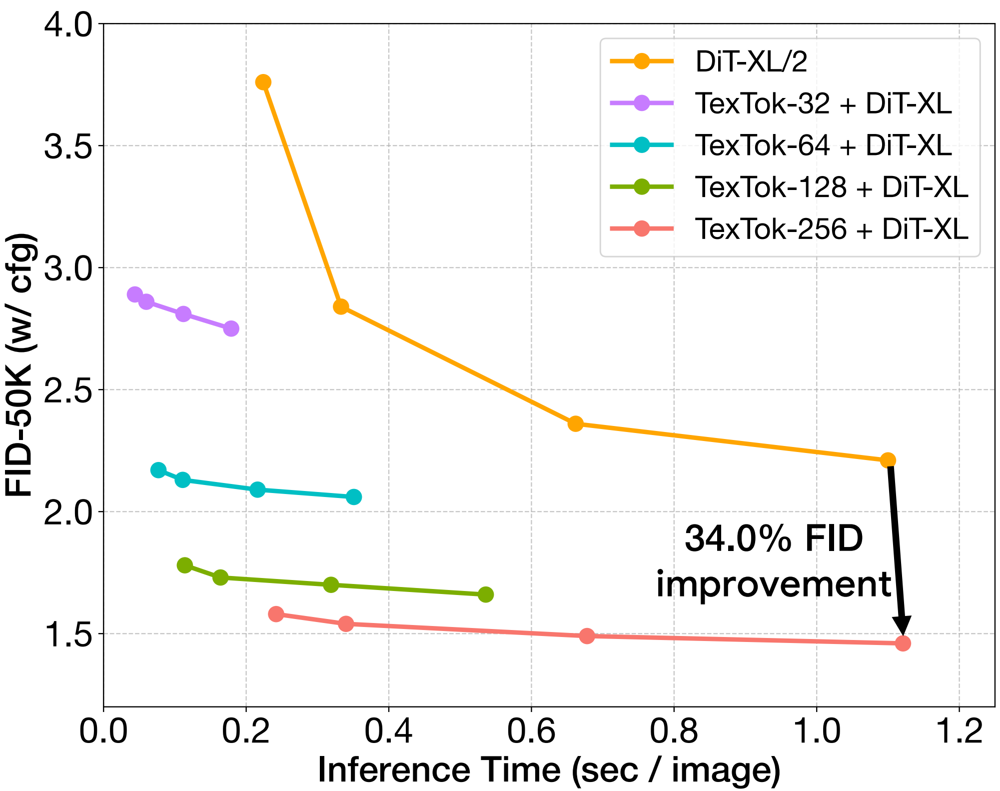
Improved Generation Speed/Quality Frontier

ImageNet 256x256



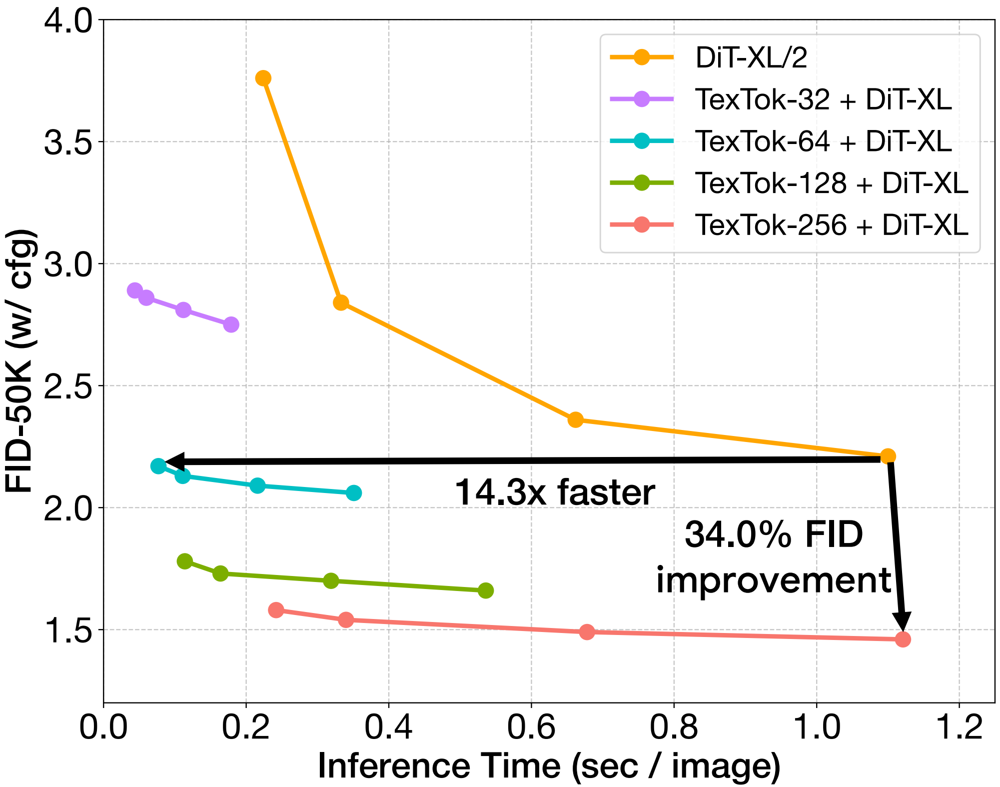
Improved Generation Speed/Quality Frontier

ImageNet 256x256

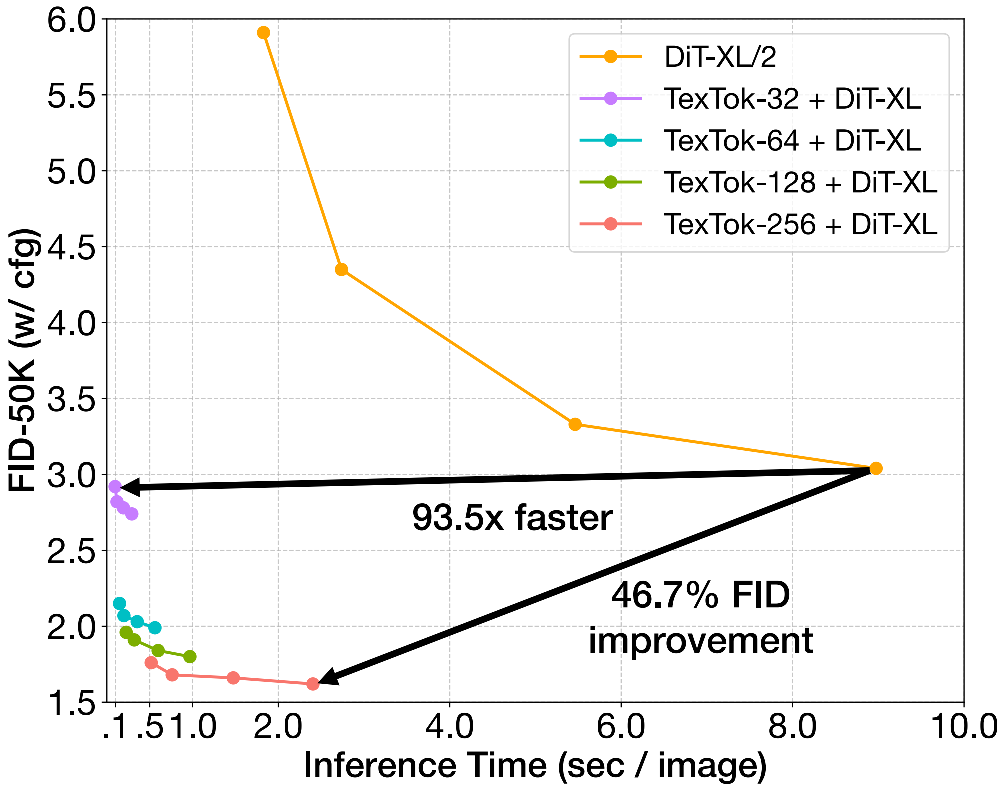


Improved Generation Speed/Quality Frontier

ImageNet 256x256



Improved Generation Speed/Quality Frontier



ImageNet 512x512

Generation Samples on ImageNet 512x512

Text-to-Image Generation

- Take the text used in generation also to tokenization
- No additional annotation cost, free performance boost

Text-to-Image Generation

Caption (Prompt): A vibrant scarlet macaw, with a striking black and white beak, perches on a weathered, grey wooden branch against a backdrop of lush green foliage. Its feathers display a gradient of red, with hints of blue and green near its tail, creating a textured and iridescent effect. The macaw's large size and bright colors make it stand out in its natural-looking environment, appearing alert and possibly watchful of its surroundings.

Text-to-Image Generation

Number of Tokens

32 64 128

Reference
Image

Baseline
(w/o text)

TexTok
(w/ text)

Text-to-Image Generation

Reference Image

Number of Tokens

32 64 128

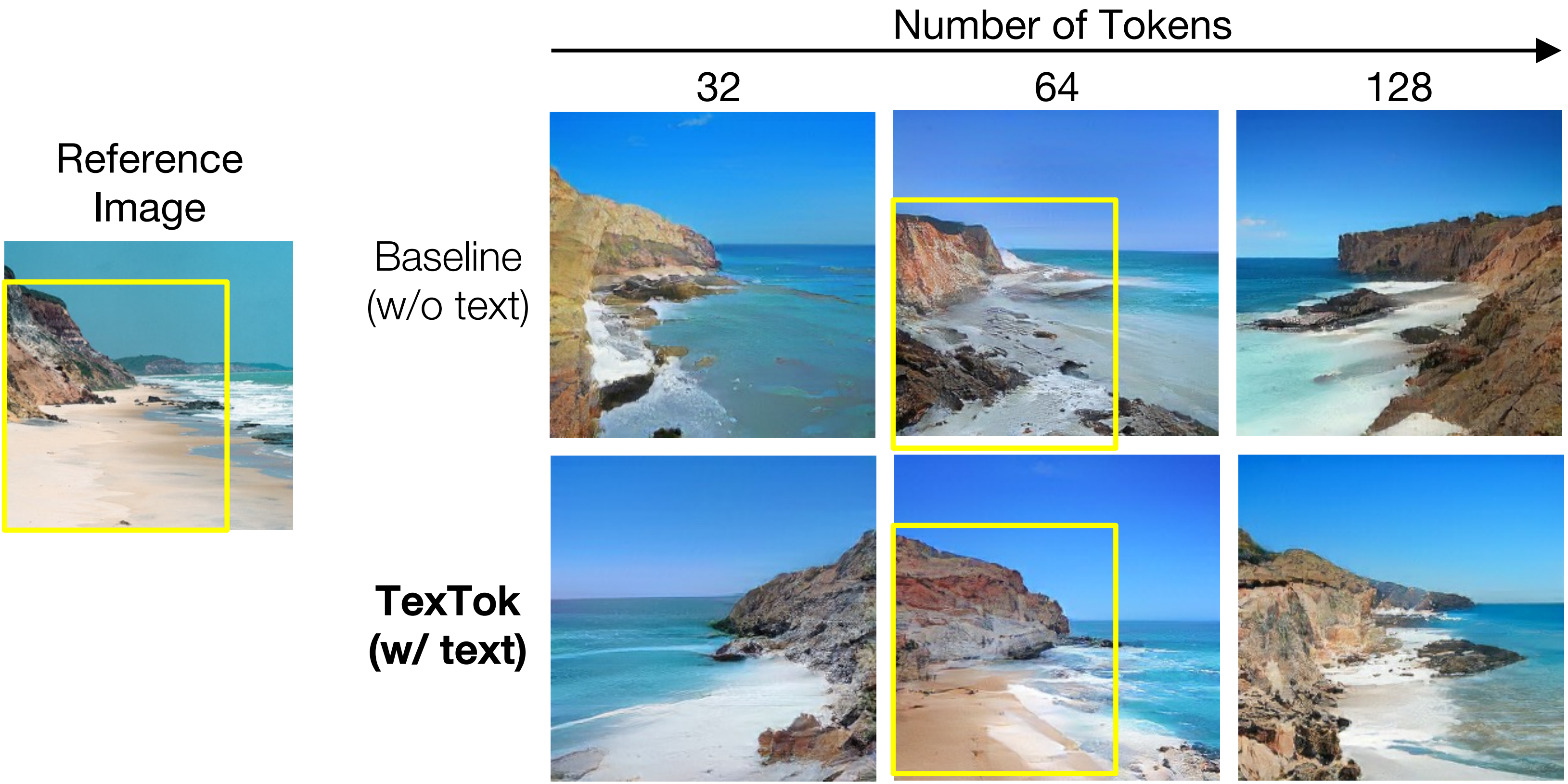
Baseline (w/o text)

TexTok (w/ text)

Text-to-Image Generation

Caption (Prompt): A towering, multi-hued cliff of red, tan, and gray rock faces a serene, turquoise ocean under a brilliant blue sky. The cliff's rough, layered texture contrasts with the smooth, white sand beach below, where gentle waves lap against dark, rocky outcrops. The expansive beach stretches alongside the cliff, forming a picturesque coastal scene under the vast, clear sky.

Text-to-Image Generation



TexTok Summary

- A tokenization framework that uses text during tokenization

- **Reconstruction**
  - better reconstruction quality
  - higher compression rate

- This leads to **generation** of
  - better generation performance
  - better computational efficiency

Check Out the Concurrent and Follow-up Work!

- TA-TiTok: https://tacju.github.io/projects/maskgen.html
- QLIP: https://nvlabs.github.io/QLIP/
- SemHiTok: https://arxiv.org/abs/2503.06764