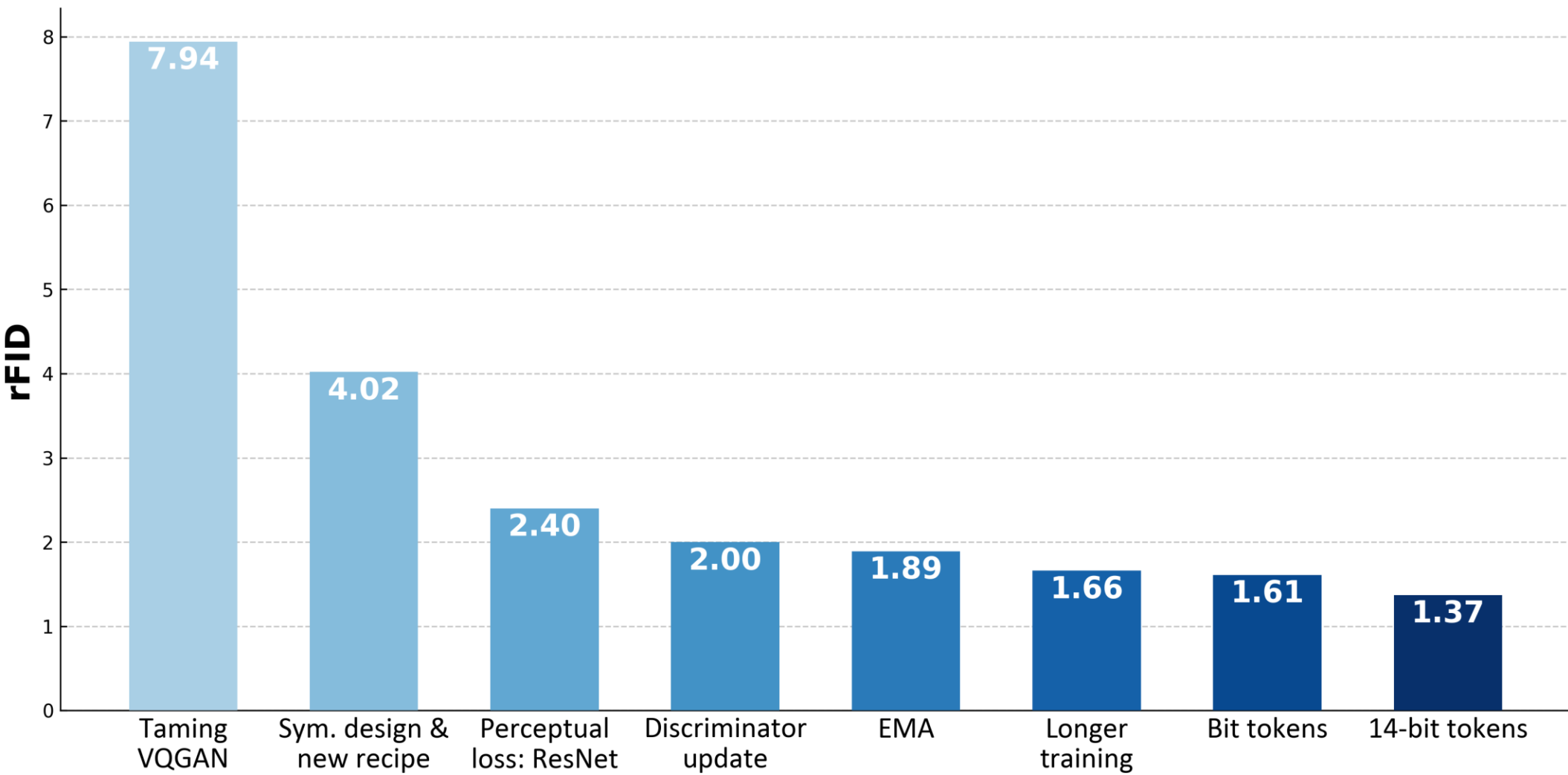


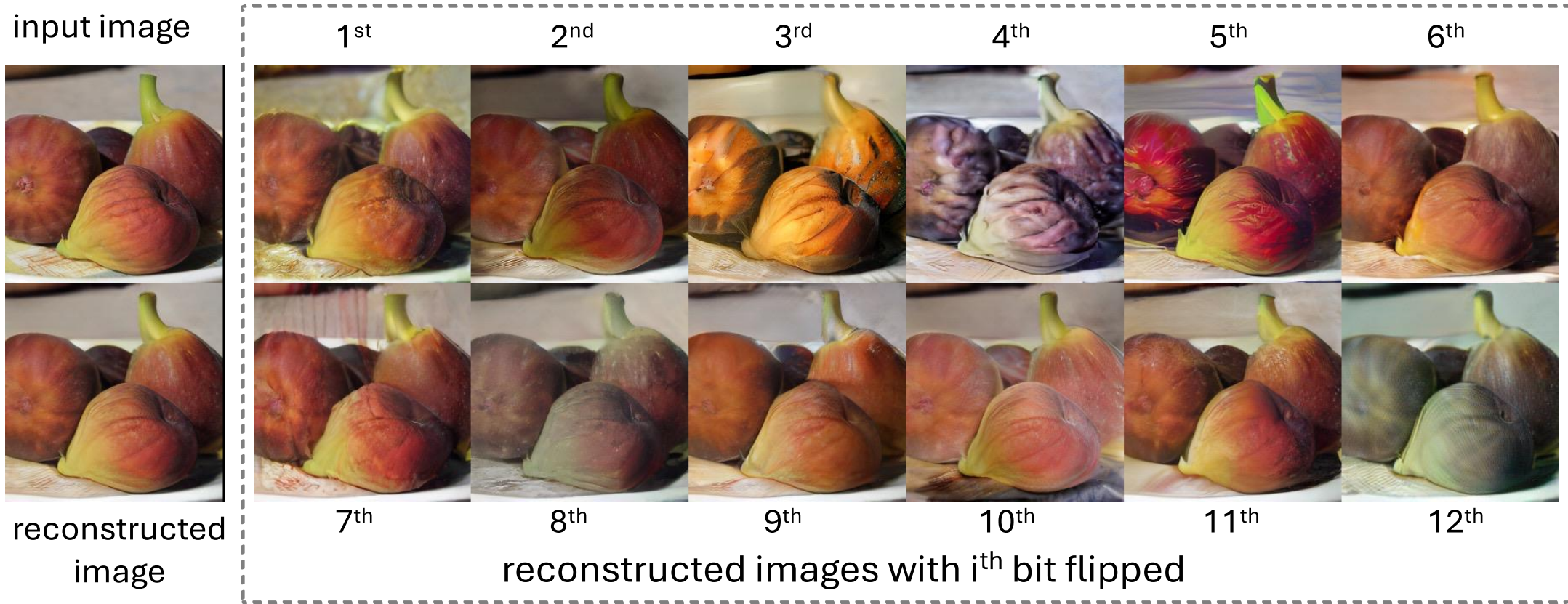
VQGAN+: A modern **tokenizer** design

Results on Imagenet 256 x 256 reconstruction.



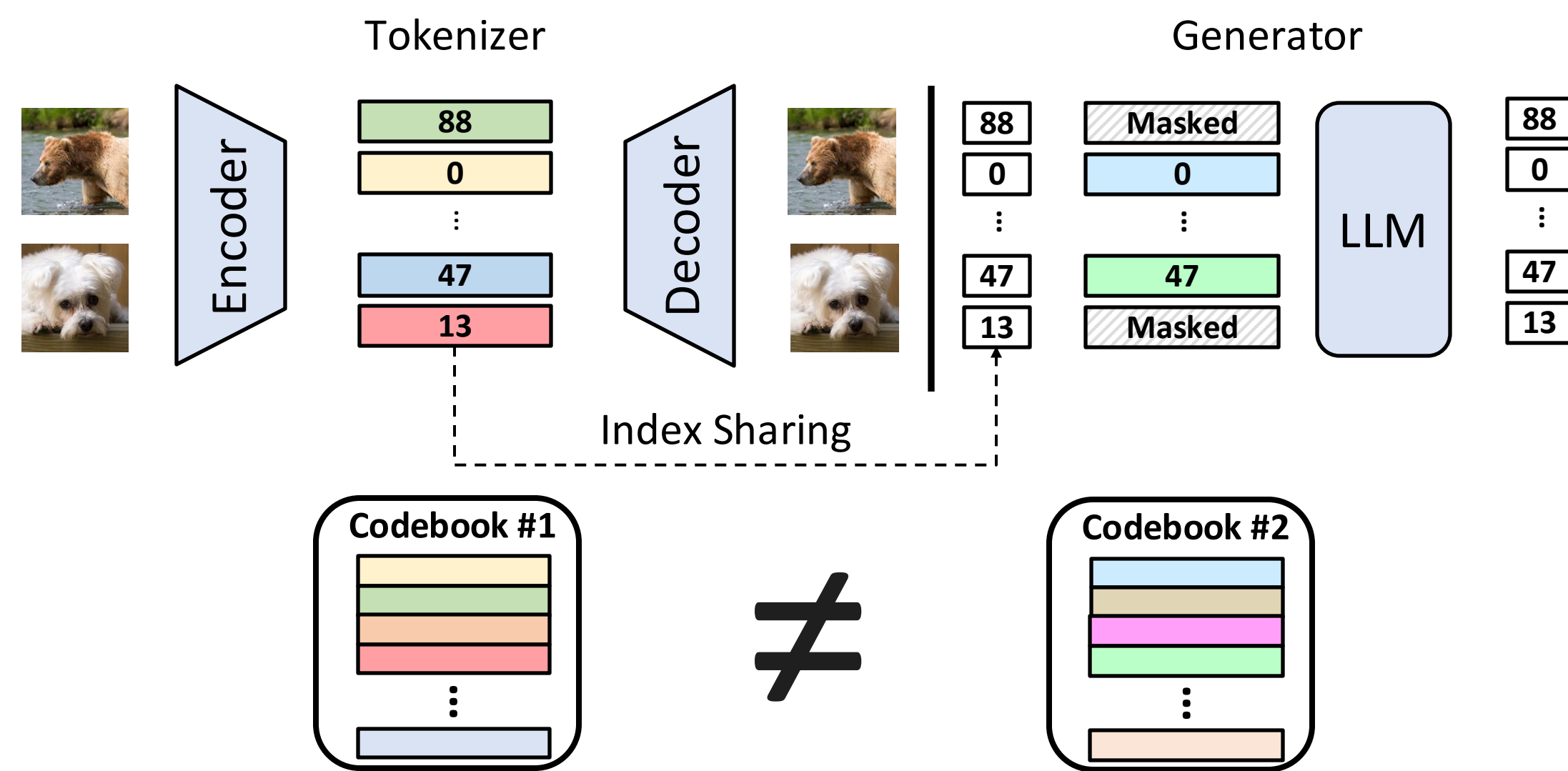
Bit tokens lead to a **semantically rich** and **disentangled** latent space

Flipping i^{th} bit in all tokens.

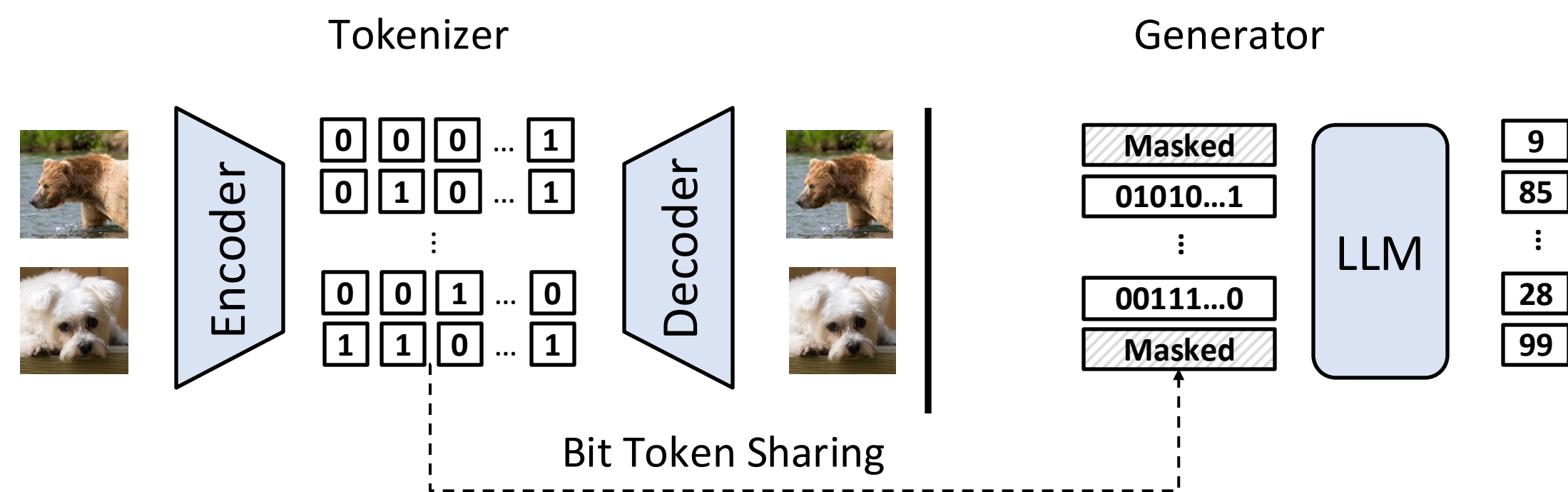


Bit tokens: A unified latent space for **tokenization** and **generation**

VQGAN and (non-) autoregressive LLMs



MaskBit



- Generator works directly on bit tokens.
- Unified representation for tokenizer and generator.
- Enables sota performance with small models.

Quantitative results

Results on Imagenet 256 x 256 class-conditional generation.

Model	Tokenizer		gFID↓	Generator		
	#tokens	codebook		IS↑	#Params	#Steps
<i>continuous latent representation</i>						
LDM-4-G (Rombach et al., 2022)	4096×3	-	3.60	247.7	400M	250
DiT-XL/2-G (Peebles & Xie, 2023)	1024×4	-	2.27	278.2	675M	250
MDT (Gao et al., 2023)	1024×4	-	1.79	283.0	676M	250
MAR-H (Li et al., 2024)*	256×16	-	1.55	303.7	943M	256
<i>discrete latent representation</i>						
Taming-VQGAN (Esser et al., 2021)	256	1024	5.88	304.8	1.4B	256
MaskGIT (Chang et al., 2022)	256	1024	4.02	355.6	177M	8
LlamaGen-3B (Sun et al., 2024)*	576	16384	2.18	263.3	3.1B	576
TiTOK-S-128 (Yu et al., 2024c)*	128	4096	1.97	281.8	287M	64
VAR (Tian et al., 2024)*	680	4096	1.92	350.2	2.0B	10
MAGVIT-v2 (Yu et al., 2024a)	256	262144	1.78	319.4	307M	64
MaskGIT (w/ our VQGAN+)	256	4096	2.12	300.8	310M	64
MaskBit (ours)	256	4096	1.65	341.8	305M	64
MaskBit (ours)	256	16384	1.62	338.7	305M	64
MaskBit (ours)	256	16384	1.52	328.6	305M	256

Ablations

Codebooks (VQ) vs embedding-free (bit tokens).

Tokenizer embedding-free	Generator embedding-free	gFID↓
✗	✗	2.12
✓	✗	1.95
✓	✓	1.82
✗	✓	2.83

High-fidelity **generation** via bit tokens

