Multimodal Latent Language Modeling with Next-Token Diffusion

Yutao Sun*[1], Hangbo Bao*[2], Wenhui Wang*[2], Zhiliang Peng*[2], Li Dong*[2], Shaohan Huang[2], Yaoyao Chang[2], Jianyong Wang[1], Furu Wei [2]

Tsinghua University[1], Microsoft Research[2]

Presenter: Yutao Sun

syt23@mails.tsinghua.edu.cn

Motivation

Unified Multimodal Generation

Tsinghua University

Intrinsic modality gap

- Text is naturally discrete tokens
- Images / audio are continuous real-world signals

Discrete tokenization bottleneck

- LM-compatible unified interface, but compression suffers from information bottleneck

Diffusion compatibility

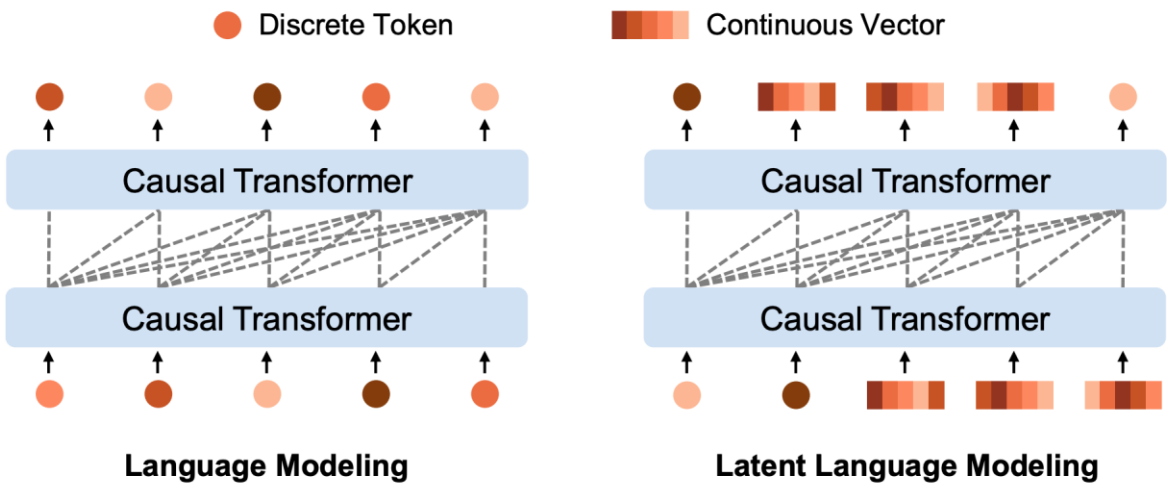
- Diffusion improves high-fidelity continuous generation
- Parallel diffusion is not naturally compatible with AR-style LMs

Framework

Continuous Auto-regressive Generation

- Consider discrete and continuous as a unified sequence
- Next token prediction with modality-aware head:

$$Decode(x_i|x_{<i}) = \begin{cases} Sample(softmax(h_i W_v), x_i \in D \\ Diffusion(h_i), x_i \in C \end{cases}$$



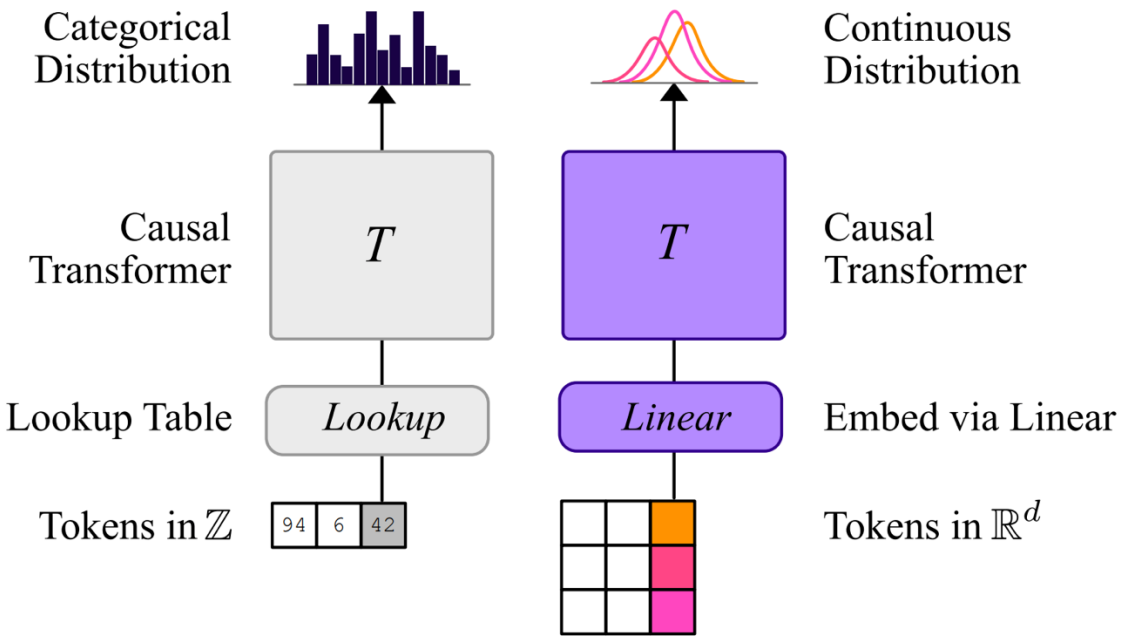
Challenge
Training-Inference Mismatch

Discrete token:

- The next token is sampled from a categorical distribution over token IDs
- The input token and target output are naturally aligned in the same discrete space

Continuous token:

- During training, the input is often the ground-truth continuous signal
- The model predicts a distribution centered around the ground truth



[1]GIVT: Generative Infinite-Vocabulary Transformers

σ-VAE

The Keystone for Next-Token Diffusion

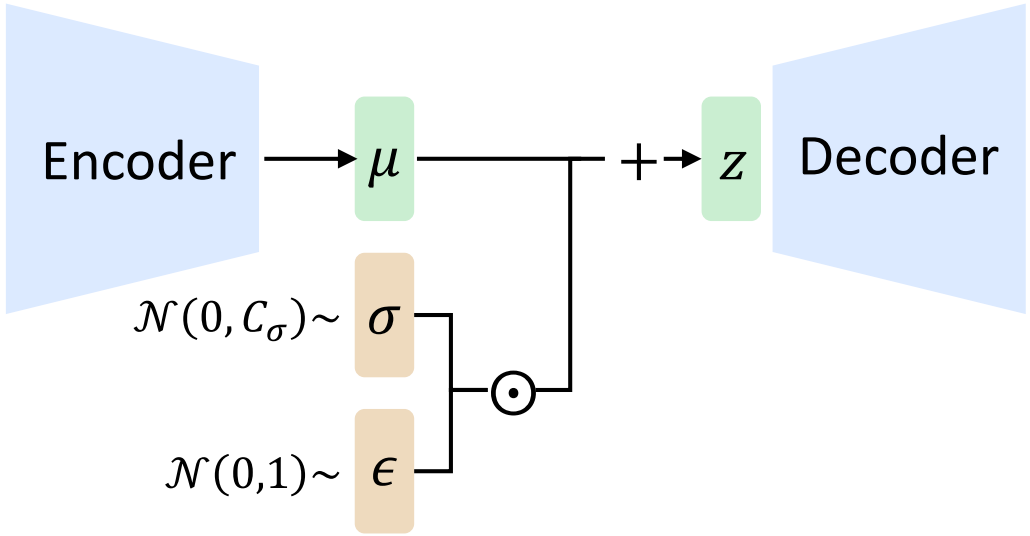
Tsinghua University

- σ-VAE enforces controllable latent variance:

$$Z = \mu + |\sigma| \cdot \epsilon,$$

$$\epsilon \in N(0,1), \sigma \in N(0, C_\sigma)$$

- Prevents collapsed latent variance in β-VAE.
- Keeping latents closer to the training distribution and improves robustness to exposure bias.



Experiment
Image Generation

Tsinghua University

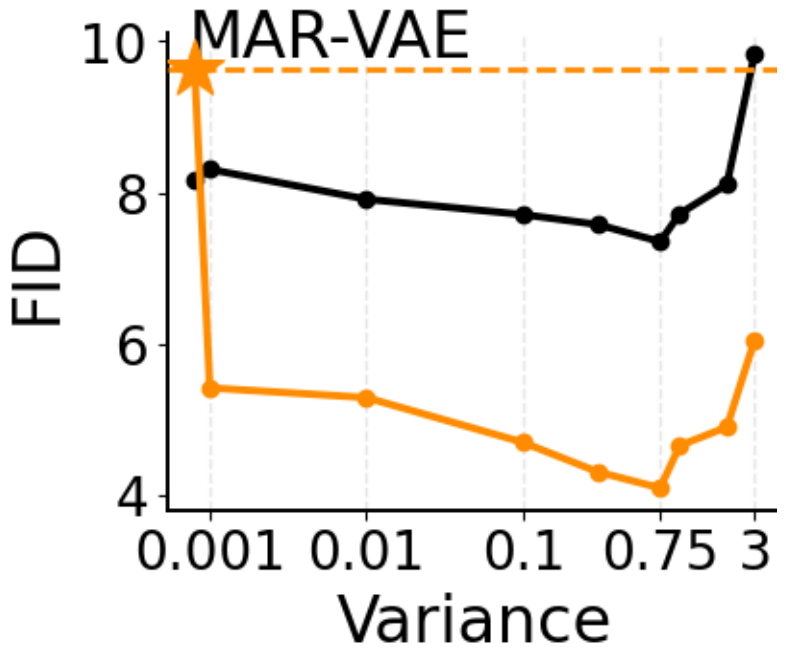
- Outperforms previous auto-regressive baselines
- Comparable with image-level DiT

Type	Model	#Params	#Epochs	FID↓	IS↑
Non-Causal-Masking Generation					
Diffusion	LDM-4 (Rombach et al., 2022)	400M	—	3.60	247.7
	DiT-XL/2 (Peebles & Xie, 2023)	675M	400	2.27	278.2
	U-ViT-H/2 (Bao et al., 2023a)	501M	400	2.29	263.9
Masked Generative	MaskGIT (Chang et al., 2022)	227M	300	4.02	355.6
	MAR-L (Li et al., 2024)	479M	800	1.78	296.0
Causal-Masking Generation					
Causal-Discrete	VQGAN (Esser et al., 2021)	1.4B	240	5.20	280.3
	ViT-VQGAN (Yu et al., 2021)	1.7B	240	3.04	227.4
	LlamaGen-XL (Sun et al., 2024a)	775M	300	2.62	244.1
	LlamaGen-XXL (Sun et al., 2024a)	1.4B	300	2.34	253.9
Causal-Continuous	GIVT-Causal-L+A (Tschannen et al., 2023)	1.67B	500	2.59	—
	LatentLM-L (This Work)	479M	400	2.24	253.8

Experiment

Effects of Tokenizer

- The tokenizers tuned for previous image-level diffusion models are ineffective for LatentLM
- LatentLM favors tokenizers with larger variances
- σ-VAE improves FID from 9.64 to 4.10

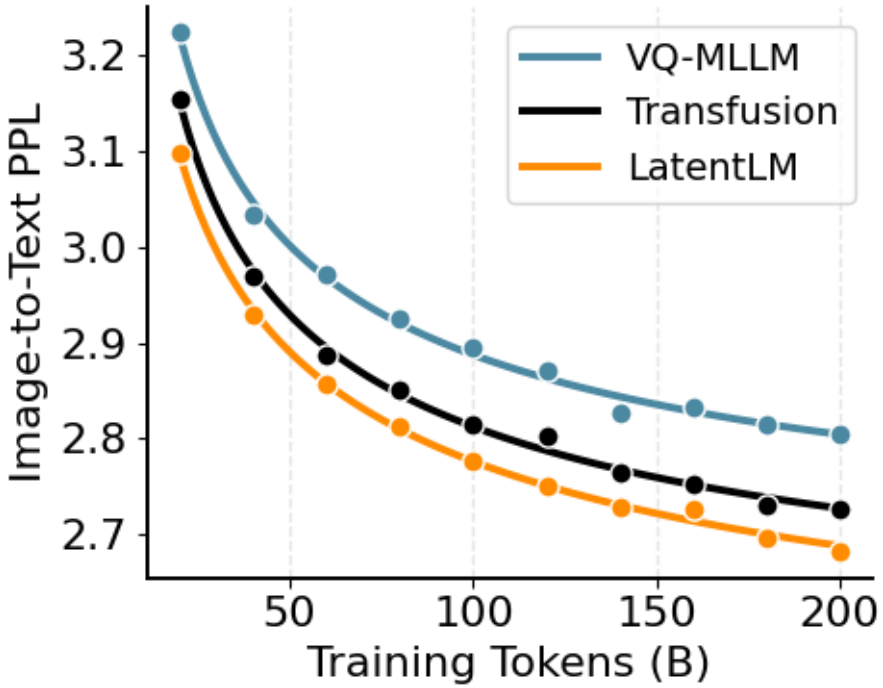


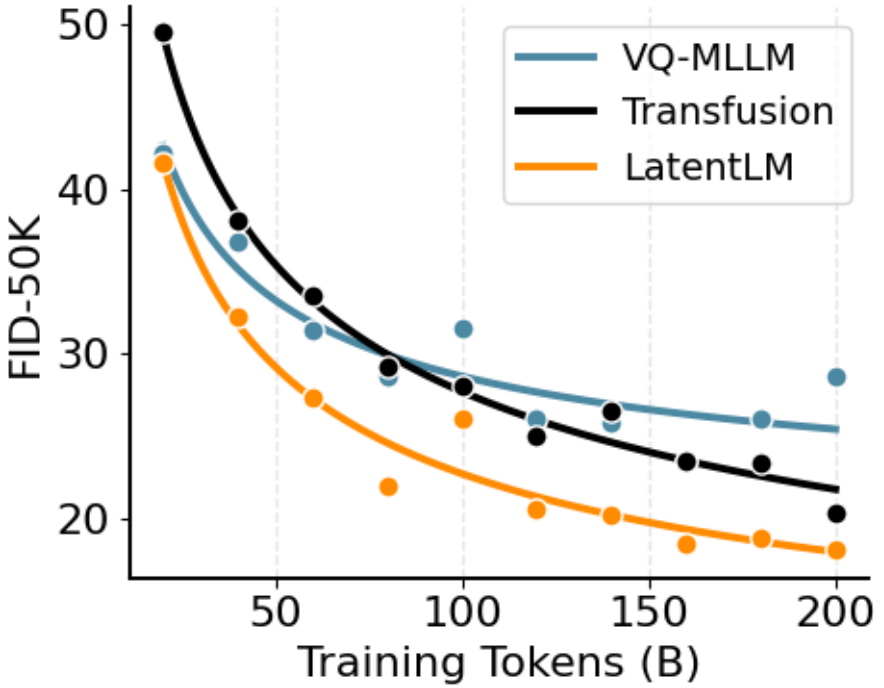
VAE	Variance	FID↓ DiT	FID↓ LatentLM
MAR (Li et al., 2024)	8e-4	**8.16**	9.64
σ-VAE	0.75	7.35	**4.10**

Experiment
Multimodal LLMs

Tsinghua University

- LatentLM provides natural interface to both understanding and generation
- Understanding: The input image of LatentLM is continuous and clean representation
- Generation: Unified auto-regressive target benefits performance





Experiment

Text-to-Speech

Tsinghua University

Higher token reduction ratio

- VQ tokenizers struggle to reduce token numbers

Fewer decoding steps

- Achieves lower inference latency with better performance

System	Frame Rate Length/s ↓	Ref Utterance as Prompt			3s Prefix as Prompt		
		SIM↑	WER-C↓	WER-H↓	SIM↑	WER-C↓	WER-H↓
Ground Truth	-	0.779	1.6	2.2	0.668	1.6	2.2
VALL-E 2 (Chen et al., 2024)	75	0.643	1.5	2.4	0.504	1.6	2.3
Voicebox (Le et al., 2023)	100	0.662	-	1.9	**0.593**	-	**2.0**
MELLE (Meng et al., 2024)	62	0.625	1.5	2.1	0.508	1.5	**2.0**
LatentLM	15	**0.697**	**1.2**	1.8	0.571	**1.4**	**2.0**
LatentLM	7.5	0.656	**1.2**	**1.7**	0.532	1.6	2.3
LatentLM	3.75	0.598	1.7	2.3	0.467	3.1	4.5

Conclusion

- **Latent Language Modeling** A causal Transformer that jointly models discrete and continuous tokens.
- **σ-VAE** A tokenizer designed for continuous auto-regressive modeling by avoiding collapsed variance.
- **Unified Performance** A single unified model achieves multimodal understanding and generation performance comparable to specialized architectures.