ICLR

NextStep-1

Toward Autoregressive Image Generation with Continuous Tokens at Scale

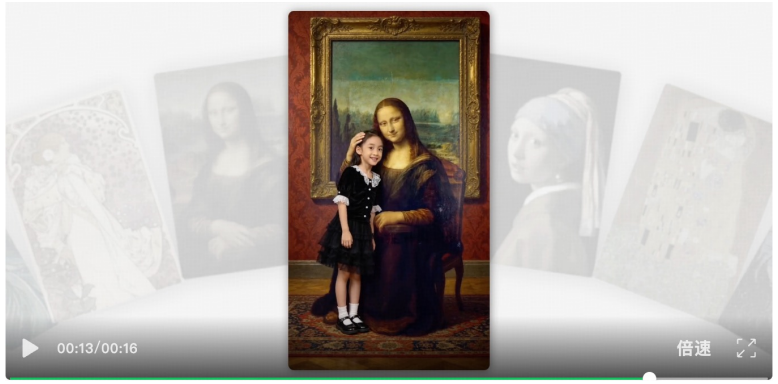
Chunrui Han*, Guopeng Li*, Jingwei Wu*, Yan Cai*, Yuang Peng*,
Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, Zheng Ge,
Shu-Tao Xia ✉, Quan Sun ✉

StepFun

* *Equal contribution (alphabetical by first name).*

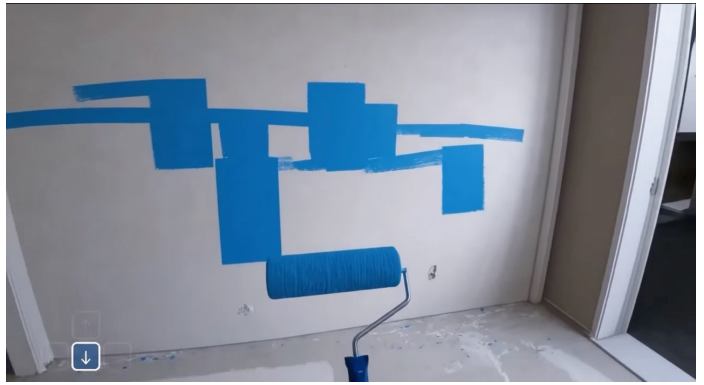
Image generation is the foundation of visual applications.

Image Generation, NextStep-1



Video Generation, Seedance

Physical Intelligence, $\pi_{0.7}$



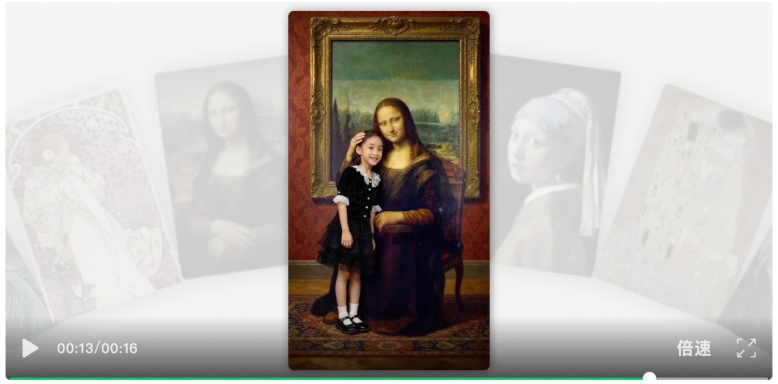
World model, Genie 3

Seedance 2.0: Advancing Video Generation for World Complexity; π_0.7: https://www.pi.website/download/pi07.pdf; Genie 3: https://deepmind.google/models/genie/

Image generation is the foundation of visual applications.

Image Generation, NextStep-1

Single High-Fidelity Image

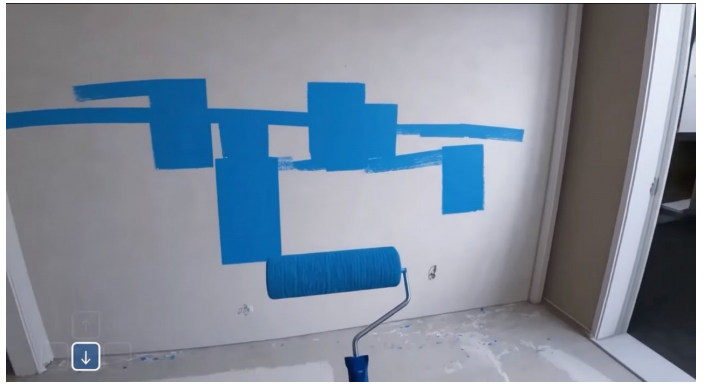


Video Generation, Seedance

Interleaved Generation and Consistency

Physical Intelligence, $\pi_{0.7}$

Next-State Spatial Prediction



World model, Genie 3

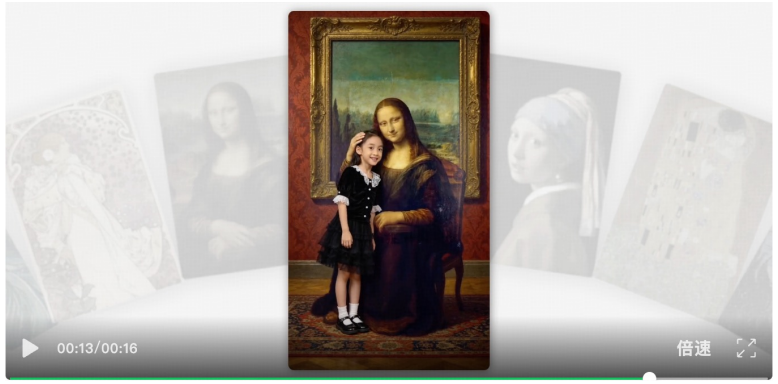
Long Memory and Context

Seedance 2.0: Advancing Video Generation for World Complexity; π_0.7: https://www.pi.website/download/pi07.pdf; Genie 3: https://deepmind.google/models/genie/

Image generation is the foundation of visual applications.

Image Generation, NextStep-1

Single High-Fidelity Image

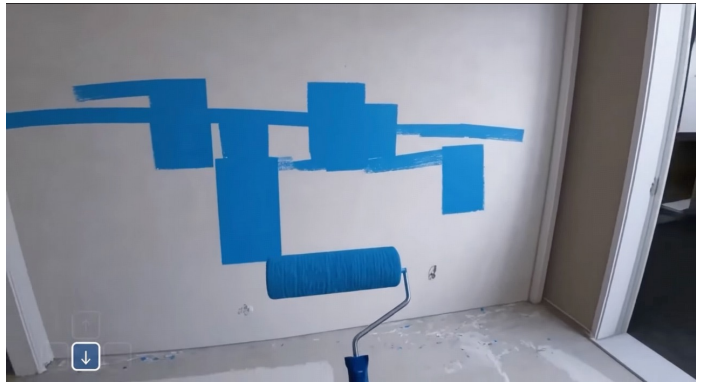


Video Generation, Seedance

Interleaved Generation and Consistency

Physical Intelligence, $\pi_{0.7}$

Next-State Spatial Prediction



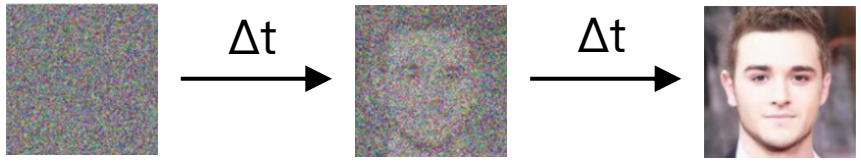
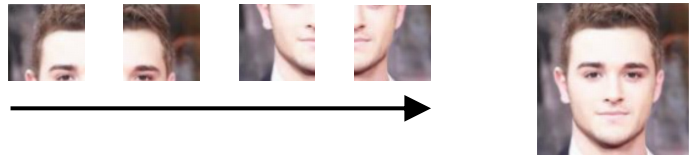
World model, Genie 3

Long Memory and Context

A single simple model for Single, Interleaved, Spatial, and Long-Context unified image generation.

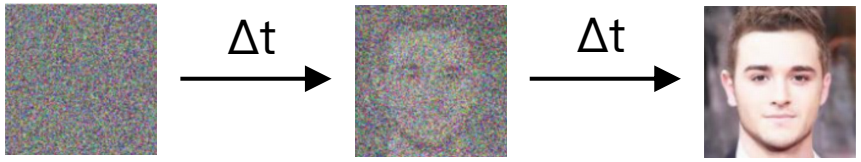
Seedance 2.0: Advancing Video Generation for World Complexity; π_0.7: https://www.pi.website/download/pi07.pdf; Genie 3: https://deepmind.google/models/genie/

Existing image models.

	Diffusion (Qwen-image, Transfusion)	Autogressive (EMU3)
		
Input	Clean + Noise image	Clean image
Tokenization	Continous	Discrete
Attention	Bidirectional	Casual
Image Modeling	Denoise	Predict
LLM Infra Compatible	Hard	Easy
Scalibity	Hard	Easy
Inference	Fast	Slow
Advantage	Painting & Fidelity	Planning & Reasoning

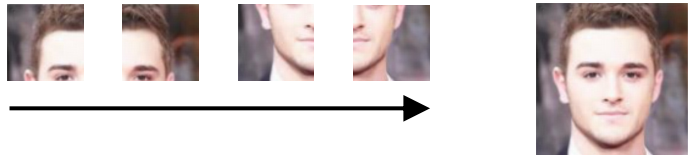
Existing image models are imperfect, each with trade-offs.

Diffusion (Qwen-image, Transfusion)



Existing diffusion models operate on continuous tokens, but are difficult to integrate with LLM infrastructure, hindering unified generation.

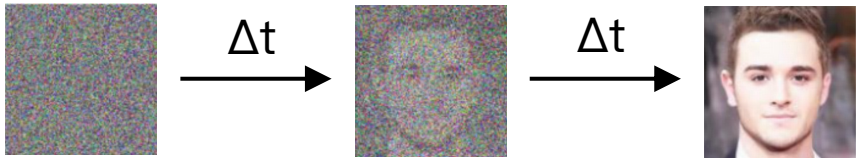
Autogressive (Chameleon, EMU3)



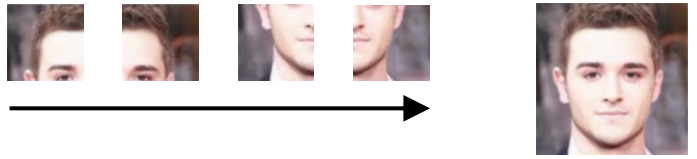
Existing autoregressive models are readily compatible with LLM infrastructure, but rely on discrete tokens that suffer from quantization loss.

Existing image models are imperfect, each with trade-offs.

Diffusion (Qwen-image, Transfusion)



Autogressive (Chameleon, EMU3)

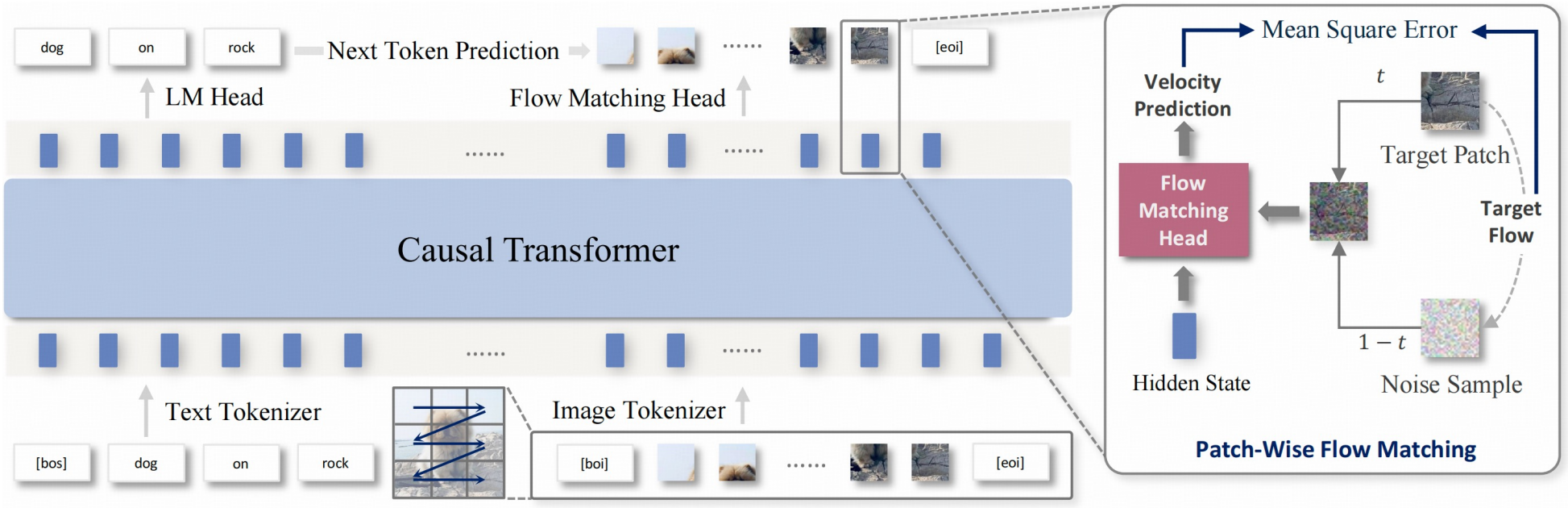


Existing diffusion models operate on continuous tokens, but are difficult to integrate with LLM infrastructure, hindering unified generation.

Existing autoregressive models are readily compatible with LLM infrastructure, but rely on discrete tokens that suffer from quantization loss.

Can we enable a standard LLM to directly process continuous tokens, achieving quantization-free unified image generation?

NextStep-1: From Discrete to Continuous Autogressive Image Generation



DESIGN

Discrete text: standard LM head with cross-entropy loss.

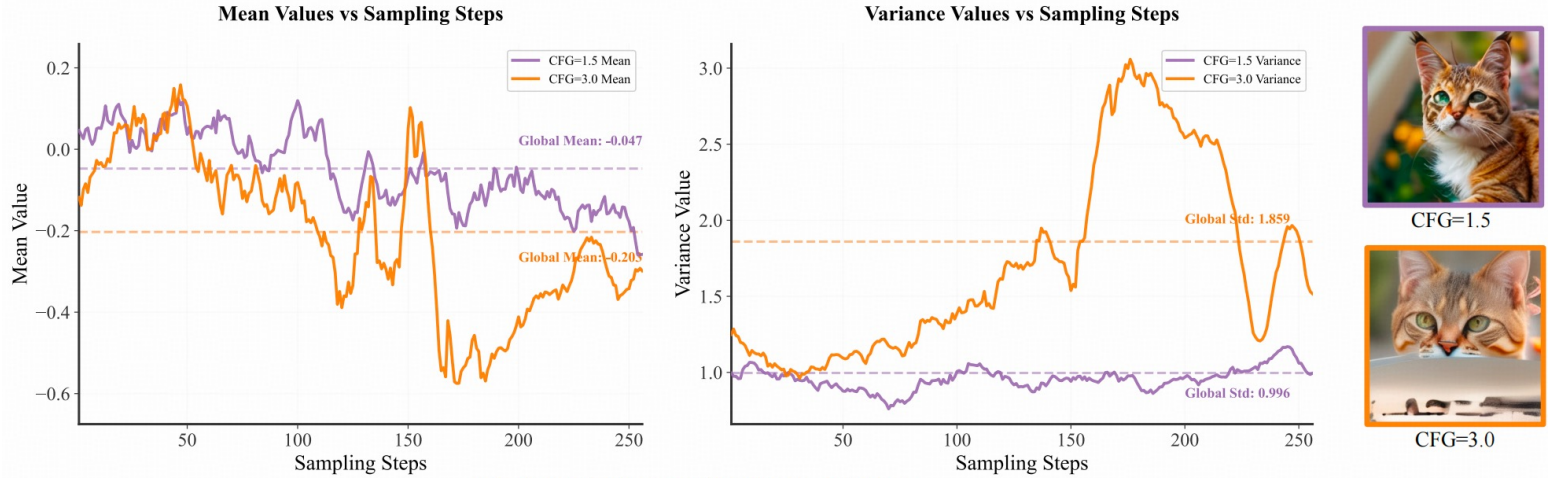
Continuous images: patch-wise flow matching head predicts visual token velocity.

Shared backbone: one causal transformer supplies contextual hidden states.

Text tokens —— **<BOI>** —— **continuous visual tokens** —— **<EOI>** —— **continue any sequence …**

Unified AR Modeling; Modality-specific decode.

Token-wise Norm: Stability Under Classifier-Free Guidance

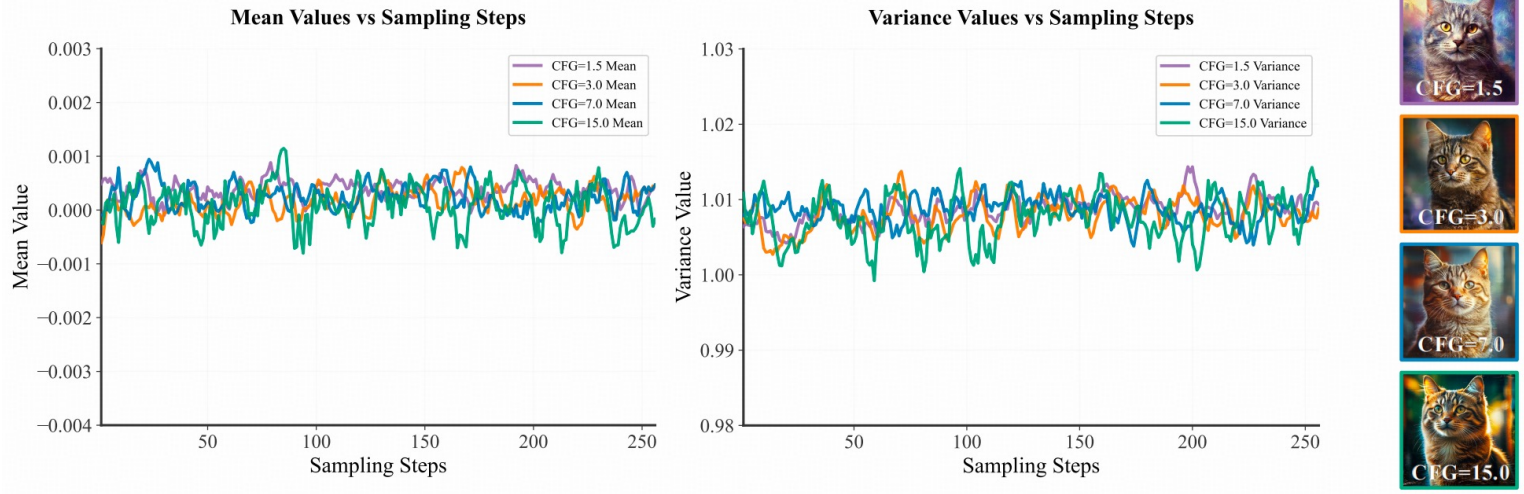


CFG=1.5

CFG=3.0

w/o channel-wise input latent normalization

w/ channel-wise input latent normalization



OBERSVATION

Moderate CFG:

Statistics stay stable.

High CFG:

Token statistics drift.

Failure Cases:

Artifacts or degraded samples.

Solution

Token-wise normalization stabilizes latent norm.

The Dilemma: Scaling the higher VAE channel dimension

RECONSTRUSTION

Original Image

SOTA f8d4 (PSNR 26.22)

Our f8d16 (PSNR 30.6)

	Batch	C4 PPL (↓)
	1M Text Tokens	10.1
Diffusion	+ 1M Image Patches	(+0.3) 10.4
Stability Modifications	1M Text Tokens	(+0.9) 11.0
LM Loss on Image Tokens	+ 1M Image Tokens	(+0.8) 11.8

rmance of the 0.76B Transfusion and Chameleon models on text
e original Llama 2 recipe.

Attention	C4 PPL (↓)	Wiki PPL (↓)	Llama Acc (↓)	CDr (↑)	MS-COC FID (↓)
Causal	10.4	6.0	51.4	12.7	61.
Bidirectional	10.4	6.0	51.7	16.0	20.
Causal	10.3	5.9	52.0	23.3	16.
Bidirectional	10.3	5.9	51.9	25.4	16.

rmance of 0.76B Transfusion models with and without intra-i
size is set at 2×2 latent pixels.

	Batch	C4 PPL (↓)
	1M Text Tokens	10.1
Diffusion	+ 1M Image Patches	(+0.3) 10.4
Stability Modifications	1M Text Tokens	(+0.9) 11.0
LM Loss on Image Tokens	+ 1M Image Tokens	(+0.8) 11.8

rmance of the 0.76B Transfusion and Chameleon models on text
e original Llama 2 recipe.

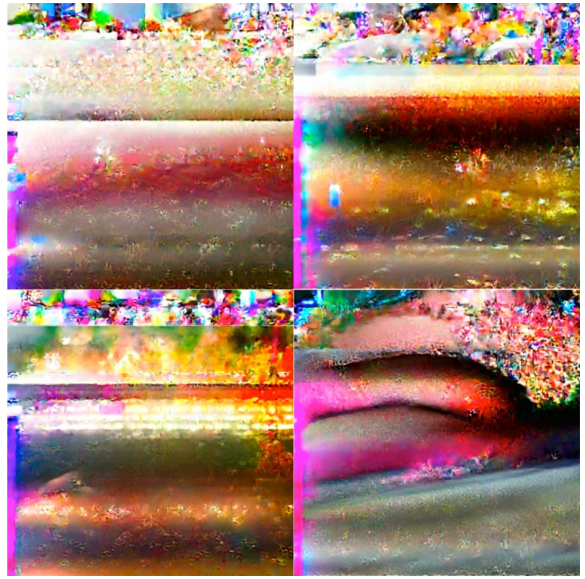
Attention	C4 PPL (↓)	Wiki PPL (↓)	Llama Acc (↓)	CDr (↑)	MS-COC FID (↓)
Causal	10.4	6.0	51.4	12.7	61.
Bidirectional	10.4	6.0	51.7	16.0	20.
Causal	10.3	5.9	52.0	23.3	16.
Bidirectional	10.3	5.9	51.9	25.4	16.

rmance of 0.76B Transfusion models with and without intra-i
size is set at 2×2 latent pixels.

GENERATION

f8d4

f8d16

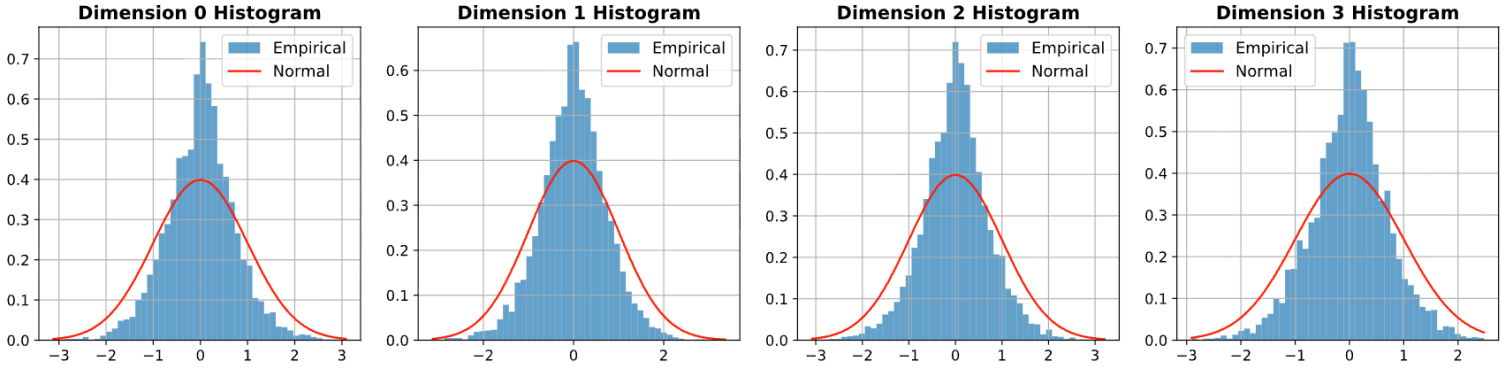


Reconstruction improves, but generation collapses.

Tokenizer Robustness Shapes the Latent Space

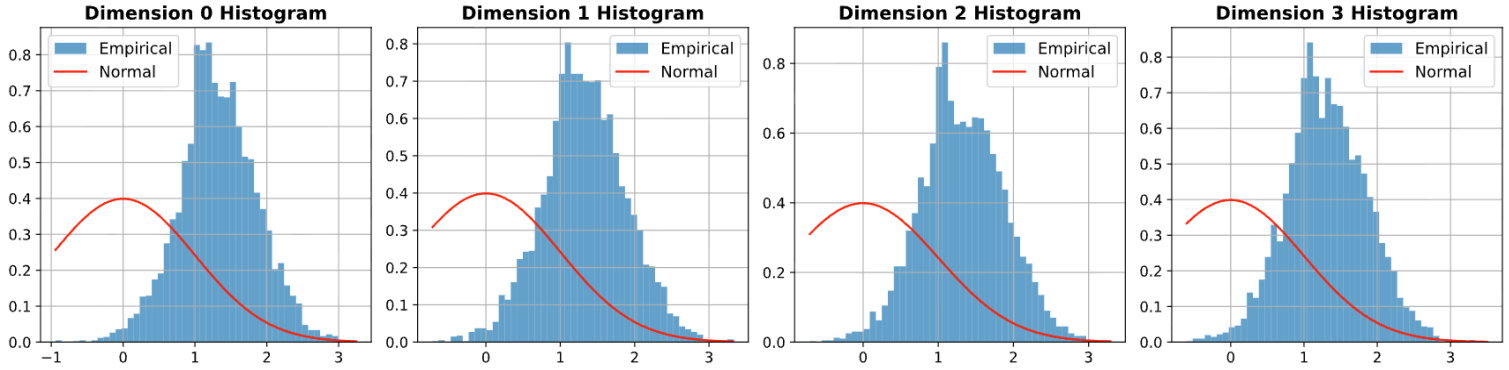
16-channel Latent Distribution

Flux.1-dev VAE



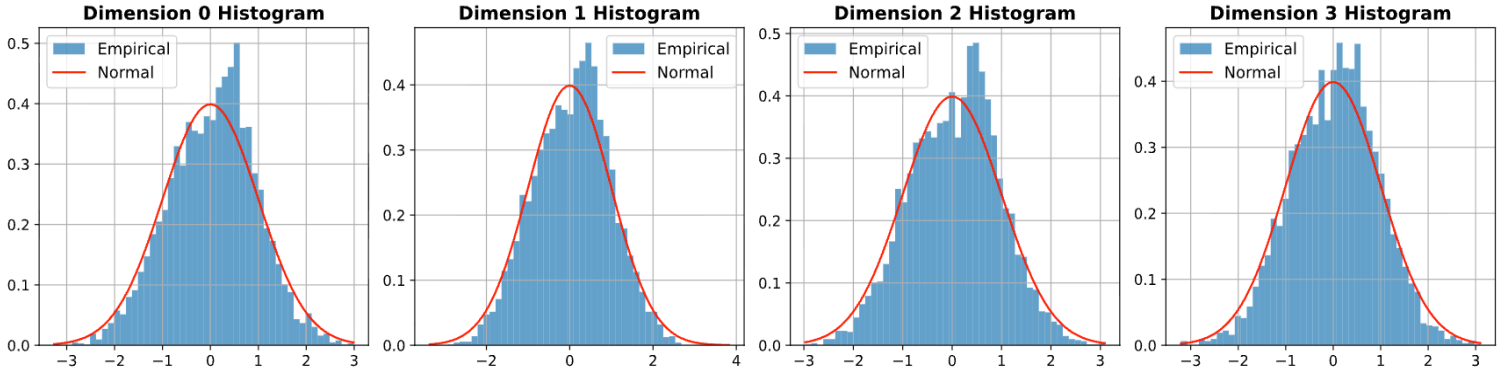
……

NextStep-1 VAE w/o Noise



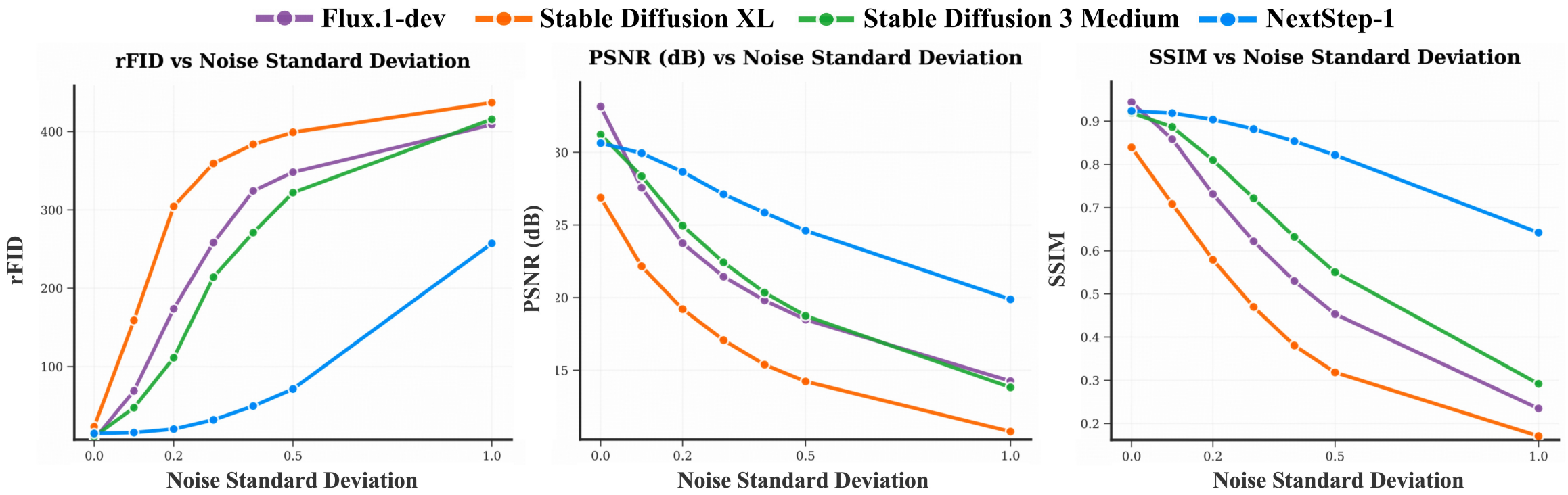
……

NextStep-1 VAE w/ Noise



……

Tokenizer Robustness Shapes the Latent Space



OBERSVATION

Optimization dilemma in image reconstruction and generation: superior reconstruction performance does not necessarily translate to better generation quality.

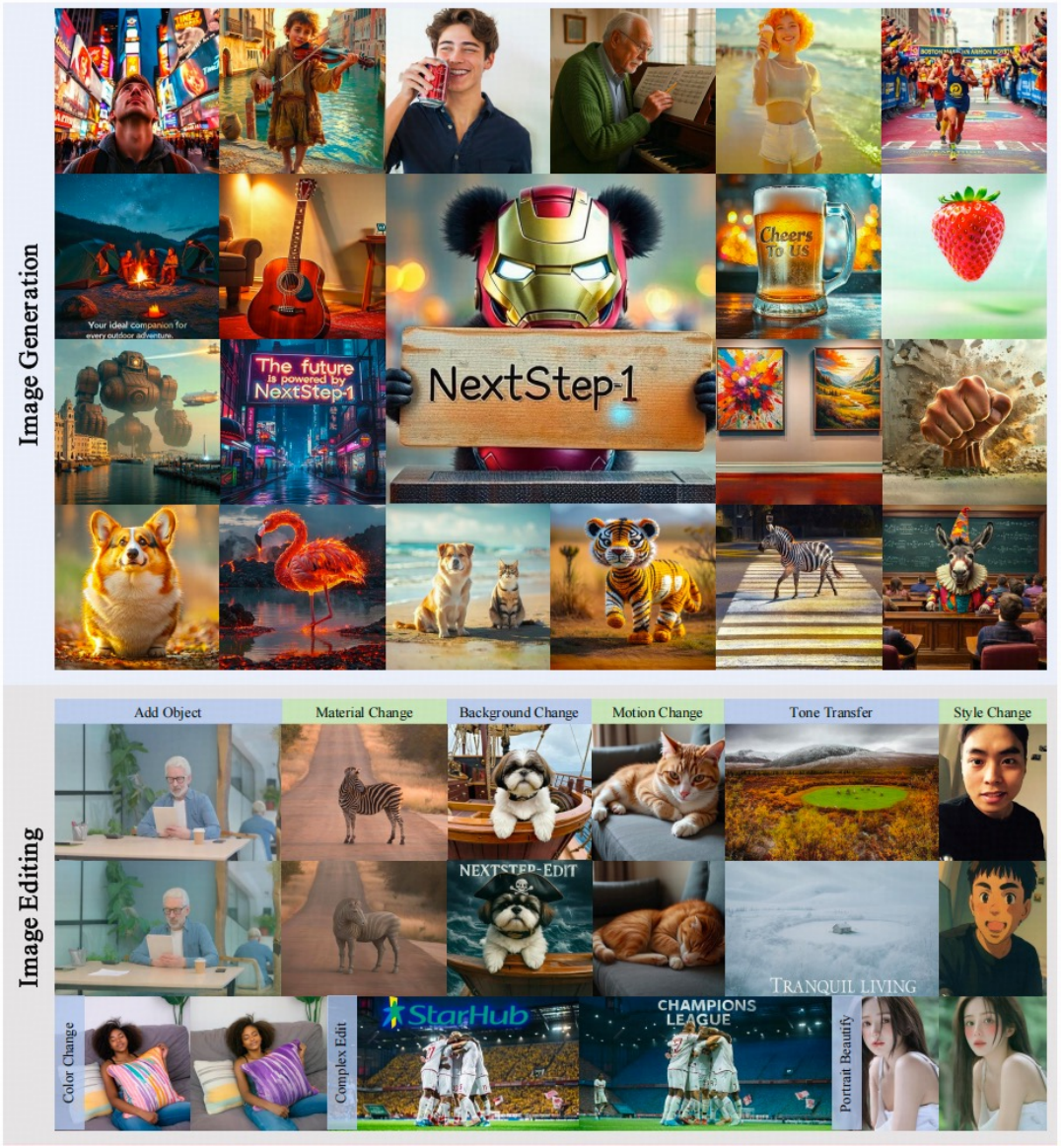
Solution

Training a VAE with noise may slightly degrade reconstruction accuracy, yet it can significantly eliminate generation collapses and improve the robustness.

Training NextStep-1

	Pre-Training			Post-Training	
	Stage1	**Stage2**	**Annealing**	**SFT**	**DPO**
Hyperparameters					
Learning Rate (Min, Max)	1×10^{-4}	1×10^{-5}	$(0, 1 \times 10^{-5})$	$(0, 1 \times 10^{-5})$	2×10^{-6}
LR Scheduler	Constant	Constant	Cosine	Cosine	Constant
Weight Decay	0.1	0.1	0.1	0.1	0.1
Loss Weight (CE : MSE)	(0.01 : 1)	(0.01 : 1)	(0.01 : 1)	(0.01 : 1)	-
Training Steps	200K	100K	20K	10K	300
Warm-up Steps	5K	5K	0	500	200
Sequence Length per Rank	16K	16K	16K	8K	-
Image Area (Min, Max)	256×256	(256×256, 512×512)	(256×256, 512×512)	(256×256, 512×512)	(256×256, 512×512)
Image Tokens (Min, Max)	256	(256, 1024)	(256, 1024)	(256, 1024)	(256, 1024)
Training Tokens	1.23T	0.61T	40B	5B	-
Data Ratio					
Text-only Corpus	0.2	0.2	0.2	0	-
Image-Text Pair Data	0.6	0.6	0.6	0.9	-
Image-to-Image Data	0.0	0.0	0.1	0.1	-
Interleaved Data	0.2	0.2	0.1	0	-

Performance of NextStep-1



Method	GenEval↑	GenAI-Bench↑		DPG-Bench↑	GEdit-EN↑
		Basic	Advanced		
Diffusion					
SD-3.5-Large	0.71	0.88	0.66	83.38	-
Flux.1-dev	0.66	0.86	0.65	83.79	-
Flux.1-Kontext-dev	-	-	-	-	6.26
Transfusion	0.63	-	-	-	-
BAGEL	**0.88**†	0.86	**0.75**†	85.07	6.52
OmniGen2	0.86†	-	-	83.57	6.41
Step1X-Edit	-	-	-	-	6.44
Qwen-Image	0.87	-	-	**88.32**	-
AutoRegressive	-				
Emu3	0.65	0.78	0.60	80.60	-
Infinity	0.79	-	-	86.60	-
Janus-Pro	0.80	0.86	0.66	84.19	-
NextStep-1	0.73†	**0.90**†	0.74†	85.28	**6.58**

- ***State-of-the-art*** **performance in autoregressive models, competing with top diffusion models**
  - GenEval → **0.73**
  - GenAI-Bench Advanced → **0.74**
  - DPG-Bench → **85.28**
- **NextStep-1-Edit:**
  - GEdit-Bench-EN (Full Set) → G_SC: **7.15** | G_PQ: **7.01** | G_O: **6.58**

+ Result is with rewriting or Self-CoT.

What governs image generation?

AR Transformer or Flow Matching Head?

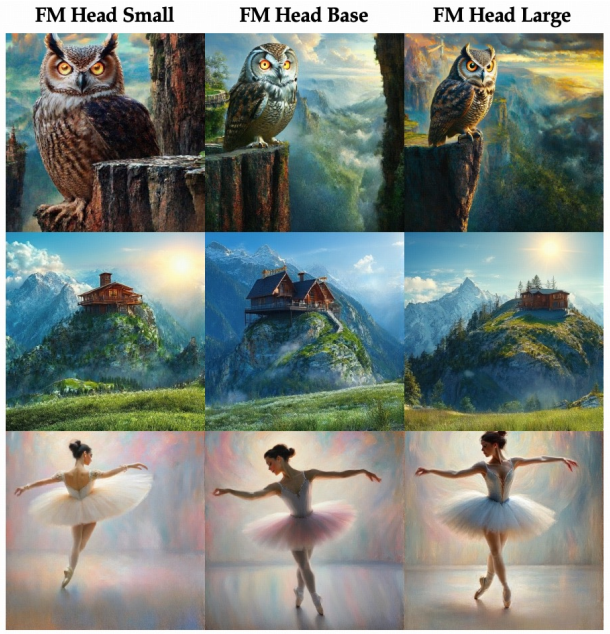


Figure 2: Images generated under different flow-matching heads.

Table 4: Configurations for different flow-matching heads.

	Layers	Hidden Size	# Parameters
FM Head Small	6	1024	40M
FM Head Base	12	1536	157M
FM Head Large	24	2048	528M

Table 5: Quantitative results for different flow-matching head configurations. All variants are finetuned from the baseline with a newly initialized head.

	GenEval	GenAI-Bench	DPG-Bench
Baseline	0.59	0.77	85.15
w/ FM Head Small	0.55	0.76	83.46
w/ FM Head Base	0.55	0.75	84.68
w/ FM Head Large	0.56	0.77	85.50

Finding:

40M/157M/528M heads behave similarly.

Interpretation:

The transformer models the conditional visual distribution, not flow-matching head.

FM head:

Lightweight continuous-token decoder.

<u>Generative logic resides in Next-Token-Prediction LLM backbone, not Flow-matching head.</u>

ICLR

NextStep-1

Toward Autoregressive Image Generation with Continuous Tokens at Scale

Chunrui Han*, Guopeng Li*, Jingwei Wu*, Yan Cai*, Yuang Peng*,
Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, Zheng Ge,
Shu-Tao Xia ✉, Quan Sun ✉

Hugging Face **Github Repo** **Arxiv Paper** **Wechat**




* Equal contribution (alphabetical by first name).