The Evolution of Visual Generative Foundation Models

Sakura

School of Earth Sciences, Zhejiang University

May 24, 2026

Outline

1. Generative Model Family
2. Latent Generative Foundation Models
3. Pixel Generative Foundation Models

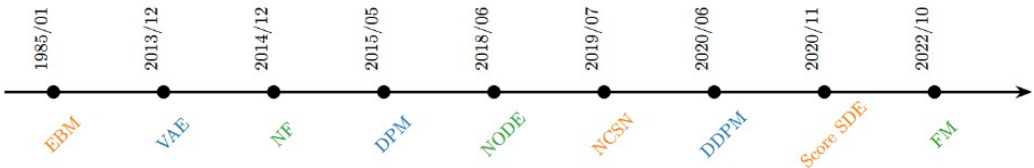
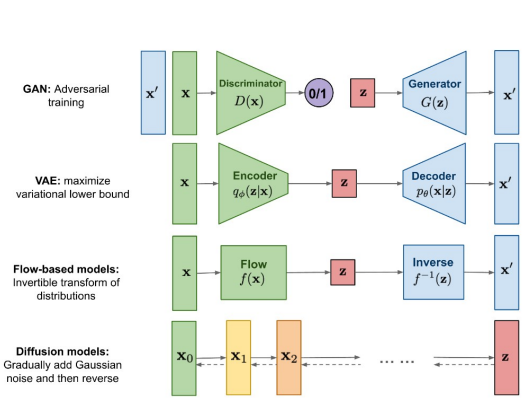


Figure: Timeline of diffusion model perspectives (Lai et al., 2025).

Lai, et al. The Principles of Diffusion Models. arXiv preprint arXiv:2510.21890. 2025.

Generative Model Family

Generative Model Family



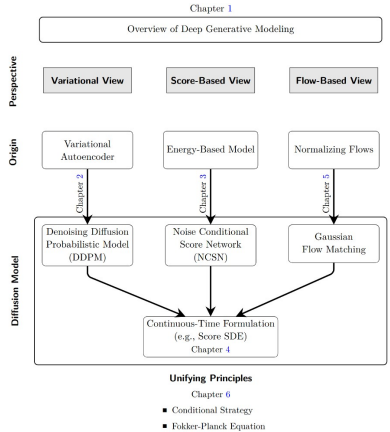
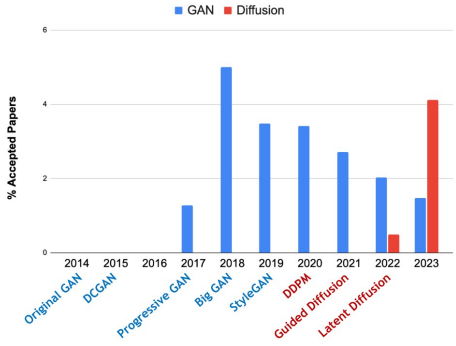


Figure: Left: Overview of different types of generative models (Weng, 2021). Right: Overview of deep generative modeling (Lai et al., 2025).

Weng, What Are Diffusion Models? lilianweng.github.io. 2021.
Lai, et al. The Principles of Diffusion Models. arXiv preprint arXiv:2510.21890. 2025.

Generative Model Family

Generally, in this talk, we mainly talk about diffusion models (e.g., score-based generative models, denoising diffusion probabilistic models, flow-based models, etc.). We also only talk about foundation models validated on class-conditional image generation on ImageNet.



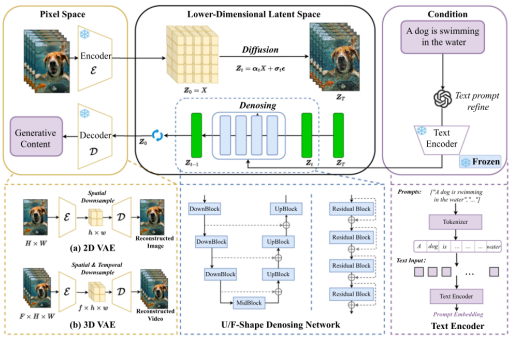


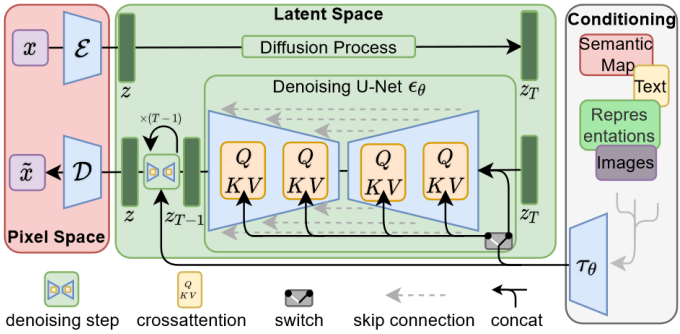
Figure: Left: GAN vs. diffusion research trend https://cvpr2022-tutorial-diffusion-models.github.io/. Right: A universal pipeline of the diffusion based models for visual content generation (Ma et al., 2025).

Ma, et al. Efficient Diffusion Models: A Comprehensive Survey From Principles to Practices. TPAMI. 2025.

Latent Generative Foundation Models

Latent Diffusion Models (LDM)

Latent Diffusion Models (LDM) (Rombach et al., 2022) first perform diffusion in a compressed latent space rather than pixel space, enabling efficient high-resolution image synthesis with flexible conditioning.



Method	FID↓	IS↑	N_{params}
SR3	11.30	—	625M
ImageBART	21.19	—	3.5B
ImageBART	7.44	—	3.5B
VQGAN+T	17.04	70.6±1.8	1.3B
VQGAN+T	5.88	**304.8±3.6**	1.3B
BigGAN-deep	6.95	203.6±2.6	340M
ADM	10.94	100.98	554M
ADM-G	4.59	186.7	608M
ADM-G, ADM-U	3.85	221.72	n/a
CDM	4.88	158.71±2.26	n/a
LDM-8	17.41	72.92±2.6	395M
LDM-8-G	8.11	190.43±2.60	506M
LDM-8	15.51	79.03±1.03	395M
LDM-8-G	7.76	209.52±4.24	506M
LDM-4	10.56	103.49±1.24	400M
LDM-4-G	**3.95**	178.22±2.43	400M
LDM-4-G	**3.60**	247.67±5.59	400M

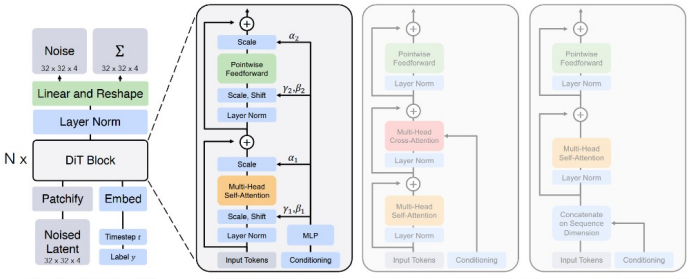
Figure: Left: LDM architecture (Rombach et al., 2022). Right: ImageNet-256 results (Rombach et al., 2022).

*In the original LDM paper, the autoencoder is a VQ-VAE from VQGAN (Esser et al., 2021). In the commercial Stable Diffusion release, it was replaced by AutoEncoderKL in a continuous latent space—the setup adopted by most subsequent latent diffusion models.

Rombach, et al. High-Resolution Image Synthesis With Latent Diffusion Models, CVPR, 2022.
Esser, et al. Taming Transformers for High-Resolution Image Synthesis, CVPR, 2021.

Diffusion Transformer (DiT)

Diffusion Transformer (DiT) (Peebles and Xie, 2023) replaces the U-Net denoiser in latent diffusion with a transformer backbone, using adaptive layer norm (adaLN-Zero) to inject timestep and class conditioning.



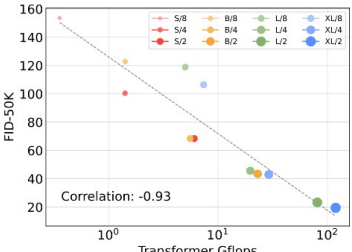
Method	FID↓	IS↑
BigGAN-deep	6.95	171.4
StyleGAN-XL	2.30	265.12
ADM	10.94	100.98
ADM-U	7.49	127.49
ADM-G	4.59	186.70
ADM-G, ADM-U	3.94	215.84
CDM	4.88	158.71
LDM-8	15.51	79.03
LDM-8-G	7.76	209.52
LDM-4	10.56	103.49
LDM-4-G (cfg=1.25)	3.95	178.22
LDM-4-G (cfg=1.50)	3.60	247.67
DiT-XL/2	9.62	121.50
DiT-XL/2-G (cfg=1.25)	3.22	201.77
DiT-XL/2-G (cfg=1.50)	**2.27**	**278.24**

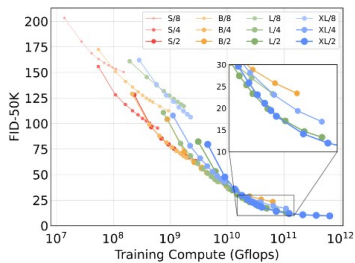
Figure: The Diffusion Transformer (DiT) architecture (Peebles and Xie, 2023). Right: ImageNet-256 results (Peebles and Xie, 2023).

Peebles, et al. Scalable Diffusion Models with Transformers, ICCV, 2023.

Diffusion Transformer (DiT): Scaling

DiT scaling follows predictable compute–quality trends: transformer Gflops correlate strongly with sample quality, and larger models use additional compute more efficiently (Peebles and Xie, 2023).





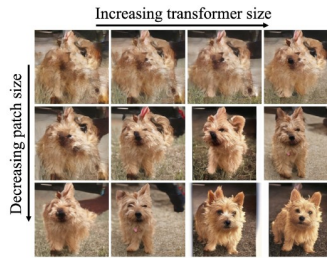


Figure: Left: Transformer Gflops are strongly correlated with FID. Larger DiT models use large compute more efficiently (Peebles and Xie, 2023). Right: Increasing transformer forward pass Gflops increases sample quality (Peebles and Xie, 2023).

Peebles, et al. Scalable Diffusion Models with Transformers, ICCV, 2023.

Scalable Interpolant Transformers (SiT)

Scalable Interpolant Transformers (SiT) (Ma et al., 2024) reuse the DiT backbone but unify flow- and diffusion-based training under stochastic interpolants, systematically ablating four design axes at matched scale.

Model	Params(M)	Training Steps	FID↓
DiT-S	33	400K	68.4
SiT-S	33	400K	**57.6**
DiT-B	130	400K	43.5
SiT-B	130	400K	**33.0**
DiT-L	458	400K	23.3
SiT-L	458	400K	**18.8**
DiT-XL	675	400K	19.5
SiT-XL	675	400K	**17.2**
DiT-XL	675	7M	9.6
SiT-XL	675	7M	**8.3**
DiT-XL (cfg=1.5)	675	7M	2.27
SiT-XL (cfg=1.5)	675	7M	**2.06**

Figure: Left: Scalable Interpolant Transformers. We systematically vary the following aspects of a generative model: time discretization, model prediction, interpolant, and sampler. ImageNet-256 results (Ma et al., 2024). Right: Selected samples from SiT-XL models trained on ImageNet at 512×512 and 256×256 resolution with cfg = 4.0, respectively (Ma et al., 2024).

Ma, et al. SiT: Exploring Flow and Diffusion-Based Generative Models with Scalable Interpolant Transformers, ECCV, 2024.

Scalable Interpolant Transformers (SiT): Ablations

SiT (Ma et al., 2024) ablates four design axes* at matched SiT-B/4 scale (400K steps, ImageNet-256). Lower FID is better.

Time discretization.

	Model	Objective	FID↓
DDPM	Noise	$\mathcal{L}_s^N$	44.2
SBDM-VP	Score	$\mathcal{L}_s$	**43.6**

Interpolant.

Interpolant	Model	Objective	FID↓
SBDM-VP	Velocity	$\mathcal{L}_v$	39.8
Linear	Velocity	$\mathcal{L}_v$	34.8
GVP	Velocity	$\mathcal{L}_v$	**34.6**

Model prediction.

Interpolant	Model	Objective	FID↓
SBDM-VP	Score	$\mathcal{L}_s$	43.6
SBDM-VP	Score	$\mathcal{L}_{s_\lambda}$	**39.1**
SBDM-VP	Velocity	$\mathcal{L}_v$	39.8

Sampler ($w_t = w_t^{\text{KL}}$).

Interpolant	Model	Objective	ODE	SDE
SBDM-VP	Velocity	$\mathcal{L}_v$	39.8	37.8
Linear	Velocity	$\mathcal{L}_v$	34.8	33.6
GVP	Velocity	$\mathcal{L}_v$	34.6	**32.9**

Figure: SiT ablations on design choices (Ma et al., 2024). Continuous training, velocity prediction with GVP, and SDE sampling with w_t^{KL} give the best FID.

*The SiT authors later develop *stochastic interpolants* as a unifying framework for flows and diffusions; interested readers may see (Albergo et al., 2025).

Ma, et al. SiT: Exploring Flow and Diffusion-Based Generative Models with Scalable Interpolant Transformers, ECCV, 2024.
Albergo, et al. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. JMLR. 2025.

Flexible Vision Transformer (FiT / FiTv2)

Flexible Vision Transformer (FiT) (Lu et al., 2024) extends the DiT-style backbone to train and infer at flexible resolutions via 2D RoPE, padded latent token sequences, and masked attention. FiTv2 (Wang et al., 2024) further improves scalability and generation quality.

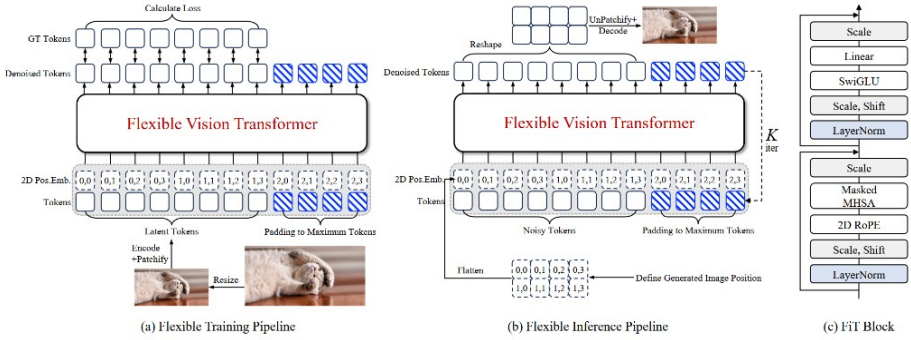


Figure: Overview of (a) flexible training pipeline, (b) flexible inference pipeline, and (c) FiT block (Lu et al., 2024; Wang et al., 2024).

Lu, et al. FiT: Flexible Vision Transformer for Diffusion Model, ICML, 2024.
Wang, et al. FiTv2: Scalable and Improved Flexible Vision Transformer for Diffusion Model, 2024.

FiT:Training-Free Resolution Extrapolation

FiT (Lu et al., 2024) extend NTK and YaRN from LLM to Vision as VisionNTK and VisionYaRN.They are training-free positional embedding interpolation approaches, especially effective in generating images with arbitrary aspect ratios.

$$\hat{f}_q(\mathbf{q}_m, h_m, w_m) = \left[e^{ih_m\Theta_h}\mathbf{q}_m \parallel e^{iw_m\Theta_w}\mathbf{q}_m\right], \quad s_h = \max\left(H_{\text{test}}/L_{\text{train}}, 1.0\right), \quad s_w = \max\left(W_{\text{test}}/L_{\text{train}}, 1.0\right),$$

with $L_{\text{train}} = \sqrt{L_{\text{max}}}$, $\Theta_h = \{\theta_d^h = b_h^{-2d/|D|}\}$, $\Theta_w = \{\theta_d^w = b_w^{-2d/|D|}\}$, and $\theta_d = b^{-2d/|D|}$ (b=10000).

VisionNTK modifies the rotary base per axis:

$$b_h = b \cdot s_h^{\frac{|D|}{|D|-2}}, \qquad b_w = b \cdot s_w^{\frac{|D|}{|D|-2}}.$$

VisionYaRN modifies the rotary frequency per axis:

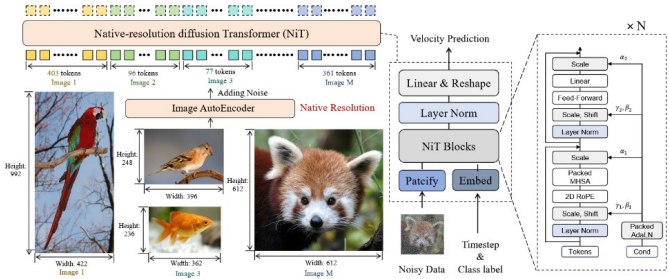
$$\theta_d^h = \left(1 - \gamma(r(d))\right)\frac{\theta_d}{s_h} + \gamma(r(d))\,\theta_d, \qquad \theta_d^w = \left(1 - \gamma(r(d))\right)\frac{\theta_d}{s_w} + \gamma(r(d))\,\theta_d,$$

where $r(d) = L_{\text{train}}/(2\pi b^{2d/|D|})$ and $\gamma(r)$ is the YaRN ramp (0 if $r<\alpha$, 1 if $r>\beta$, else linear). At aspect ratio $1{:}1$, VisionNTK/VisionYaRN reduce to vanilla NTK/YaRN.

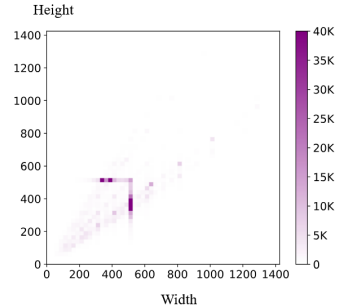
Lu, et al. FiT: Flexible Vision Transformer for Diffusion Model, ICML, 2024.

Native-resolution Image Synthesis (NiT)

Native-resolution Image Synthesis (NiT) (Wang et al., 2025) further extends FiT/FiTv2 which takes noisy latent representations, tokenizes them into variable-length sequences based on the original image resolution without padding. Flash-Attention 2 (Dao, 2023) is used to natively process heterogeneous, unpadded token sequences by cumulative sequence lengths using the memory tiling strategy.



Architecture Design of Native Resolution Diffusion Transformer (NiT) (Wang et al., 2025).



Data distribution of ImageNet.

Wang, et al. Native-Resolution Image Synthesis, NeurIPS, 2025.
Dao, FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, 2023.

Native-resolution Image Synthesis (NiT)

Method	Param	Res	Token	FID↓					
				256	512	768	1024	1536	2048
DiT-XL/2 †	675M	256/512	1428B	2.27	3.04	—	—	—	—
SiT-XL/2 †	675M	256/512	1428B	2.06	2.62	—	—	—	—
FlowDCN †	675M	256/512	158B	2.00	2.44	9.82	18.64	41.17	69.88
FiTv2-XL †	671M	256/512	237B	2.26	2.62	190.69	281.55	×	×
SiT-REPA †	675M	256/512	525B	**1.42**	2.08	274.63	286.79	×	×
PixNerd †	700M	256/512	—	2.15	2.84	—	—	—	—
EDM2-L	777M	512	472B	—	1.88	9.02	40.74	105.57	172.30
EDM2-XXL	1.5B	512	472B	—	1.81	—	—	—	—
NiT-XL	675M	Native	131B	2.16	1.57	—	—	—	—
NiT-XL	675M	Native	197B	2.03	**1.45**	**4.05**	**4.52**	**6.51**	**24.76**

Figure: Left: Benchmarking on ImageNet (Wang et al., 2025). Token: total training token budget (sum of latent tokens across all training iterations, analogous to LLM token budgets). †Independent model per benchmark. ×: failed to generalize. Right: Native-resolution image synthesis on ImageNet (Wang et al., 2025).

Wang, et al. Native-Resolution Image Synthesis, NeurIPS, 2025.

DiM & DiG: SSM as Generative Backbone

Besides standard DiT and their variants, some people also explore using SSM as generative backbone; see DiM and DiG, which also achieve competitive performance on ImageNet compared to DiT with efficiency.

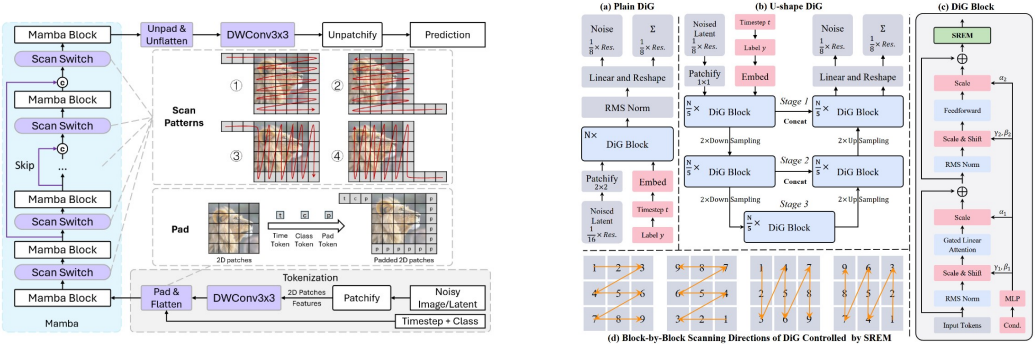
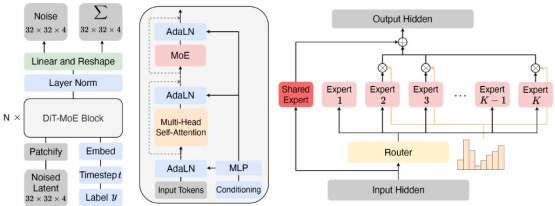


Figure: Left: DiM architecture (Teng et al., 2024). Right: The overview of the proposed DiG models (Zhu et al., 2025).

Teng, et al. DiM: Diffusion Mamba for Efficient High-Resolution Image Synthesis, 2024.
Zhu, et al. DiG: Scalable and Efficient Diffusion Models with Gated Linear Attention, CVPR, 2025.

DiT-MoE: Scaling Diffusion Transformers with Mixture-of-Experts

DiT-MoE (Fei et al., 2024) scales diffusion transformers to 16 billion parameters by replacing dense feed-forward layers with sparse Mixture-of-Experts (MoE) blocks. DiT-MoE-G/2 achieves FID 1.8 on ImageNet-512.



Model	Total	Act.	L	D	n	Gflops
S/2-8E2A	199M	71M	12	384	6	15.43
S/2-16E2A	369M	71M	12	384	6	15.44
B/2-8E2A	795M	286M	12	768	12	61.68
L/2-8E2A	2.8B	1.0B	24	1024	16	219.26
XL/2-8E2A	4.1B	1.5B	28	1152	16	323.74
G/2-16E2A	16.5B	3.1B	40	1408	16	690.94

Figure: Left: Overview of the DiT-MoE architecture (Fei et al., 2024). Right: DiT-MoE model configurations at ImageNet-256. For reference, DiT-XL/2 uses 119 Gflops and ADM uses 1120 Gflops.

*Training additionally uses synthetic datasets generated by Flux and SD3.

Fei, et al. Scaling Diffusion Transformers to 16 Billion Parameters, 2024.

Pixel Generative Foundation Models

Pixel Generative Foundation Models