An Introduction to Visual Generation

Sakura

School of Earth Sciences, Zhejiang University

May 22, 2026

Preliminary: Deep Generative Models

In early days, people from different groups explored visual generative models via denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) and score-based generative models (NCSN) (Song and Ermon, 2019). Later on, they were interpreted under the same framework called Score-SDE (Song, Sohl-Dickstein, et al., 2021). After that, novel frameworks such as Flow Matching (Lipman et al., 2023) emerged, which are out of the scope of Score-SDE.

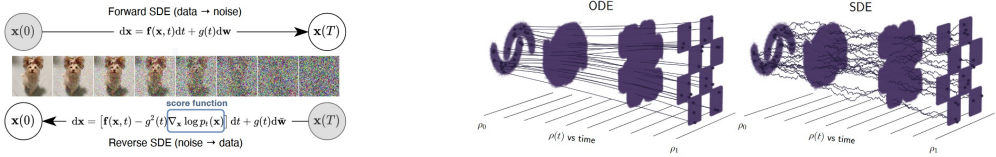
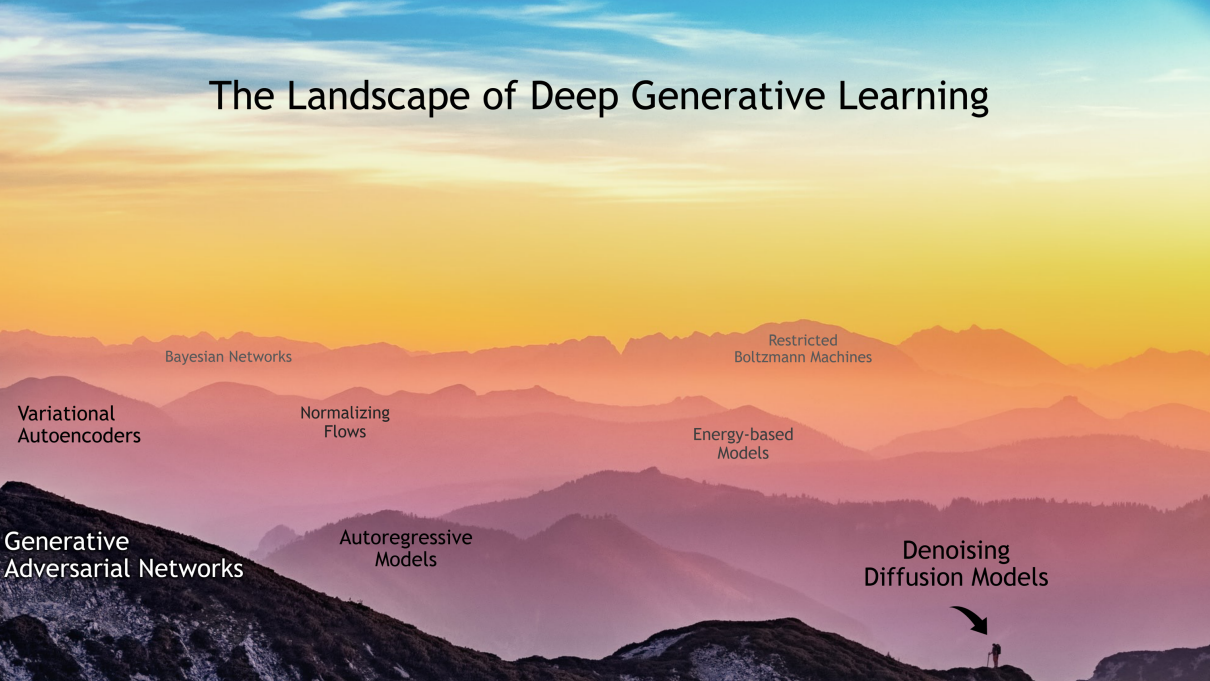


Figure: Left: An illustration of deep generative models (Song, Sohl-Dickstein, et al., 2021). Right: The stochastic interpolant paradigm (Albergo et al., 2025).

Nowadays, a promising unified framework called stochastic interpolants (Albergo et al., 2025) unifies most generative modeling frameworks. However, there is still a lack of common sense to name all these model groups with a unified category. Hence, for simplicity and clarity, we refer to them as **deep generative models** all through this presentation.

Ho, et al. Denoising Diffusion Probabilistic Models, NeurIPS, 2020.
Song, et al. Generative Modeling by Estimating Gradients of the Data Distribution, NeurIPS, 2019.
Song, Sohl-Dickstein, et al. Score-Based Generative Modeling through Stochastic Differential Equations, ICLR, 2021.
Lipman, et al. Flow Matching for Generative Modeling, ICLR, 2023.
Albergo, et al. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. TMLR. 2025.


The Landscape of Deep Generative Learning
Bayesian Networks
Restricted
Boltzmann Machines
Variational
Autoencoders
Normalizing
Flows
Energy-based
Models
Generative
Adversarial Networks
Autoregressive
Models
Denoising
Diffusion Models

Outline

1. Unconditional/Class-Conditional Image Generation
2. Text-to-Image Generation
3. Controllable Image Generation
4. Image Editing
5. Image-to-Image Translation
6. Video Generation
7. Advanced Applications
    - Inversion Task
    - Representation Extractor
    - Generalist Vision Learner
    - 3D/4D Generation & World Modeling
    - etc.

Preliminary: One Training Objective

> *"There is only one precise way of presenting the laws, and that is by means of differential equations. They have the advantage of being fundamental and, so far as we know, precise."*
> — Richard P. Feynman

Many deep generative models share one training recipe (Lai et al., 2025): fit a network to a noisy version of the data, weighted over time.

$$\mathcal{L}(\phi) := \mathbb{E}_{\mathbf{x}_0,\,\boldsymbol{\epsilon}}\left[\underbrace{\mathbb{E}_{p_{\text{time}}(t)}}_{\text{time distribution}}\left[\underbrace{\omega(t)}_{\text{time weighting}}\underbrace{\left\|\mathsf{NN}_\phi\big(\mathbf{x}_t, t\big) - \big(A_t\mathbf{x}_0 + B_t\boldsymbol{\epsilon}\big)\right\|_2^2}_{\text{MSE part}}\right]\right]$$

(A) Noise schedule in the forward process of $\mathbf{x}_t$ via α_t and σ_t.

(B) Prediction types of NN_ϕ and regression targets $\big(A_t\mathbf{x}_0 + B_t\boldsymbol{\epsilon}\big)$.

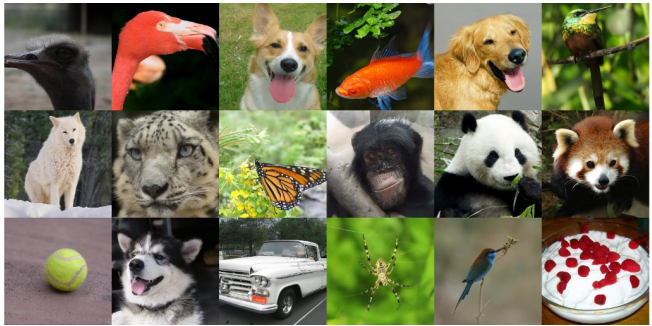
(C) Time-weighting function $\omega(\cdot) : [0, T] \to \mathbb{R}_{\geq 0}$.

(D) Time distribution p_{time}.

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

Unconditional Image Generation

This section covers $p(\mathbf{x})$ and class-conditional $p(\mathbf{x} \mid c)$, where c is a discrete class label (e.g., an ImageNet category)—not a text prompt or visual control signal.



Res.	Model	FID	sFID	Prec	Rec
128^2	BigGAN-deep	6.02	7.18	**0.86**	0.35
	ADM-G	**2.97**	**5.09**	0.78	0.59
	ADM-U	5.91	**5.09**	0.70	**0.65**
256^2	BigGAN-deep	6.95	7.36	**0.87**	0.28
	ADM-G	**4.59**	**5.25**	0.82	**0.52**
	ADM-U	10.94	6.02	0.69	0.63
512^2	BigGAN-deep	8.43	8.13	**0.88**	0.29
	ADM-G	**7.72**	**6.57**	0.87	0.42
	ADM-U	23.24	10.19	0.73	**0.60**

Figure: Selected samples from the best ImageNet 512×512 model (FID 3.85) from ADM (Dhariwal and Nichol, 2021). Right: BigGAN-deep (Brock et al., 2019) vs. ADM-G (classifier guidance ADM) vs. ADM-U (unconditional ADM).

Dhariwal, et al. Diffusion Models Beat GANs on Image Synthesis, NeurIPS, 2021.
Brock, et al. Large Scale GAN Training for High Fidelity Natural Image Synthesis, ICLR, 2019.

Text-to-Image Generation

Synthesize images from natural language descriptions. Latent Diffusion Models (LDM; Stable Diffusion) (Rombach et al., 2022) led the era of latent modeling in research and AIGC in business.

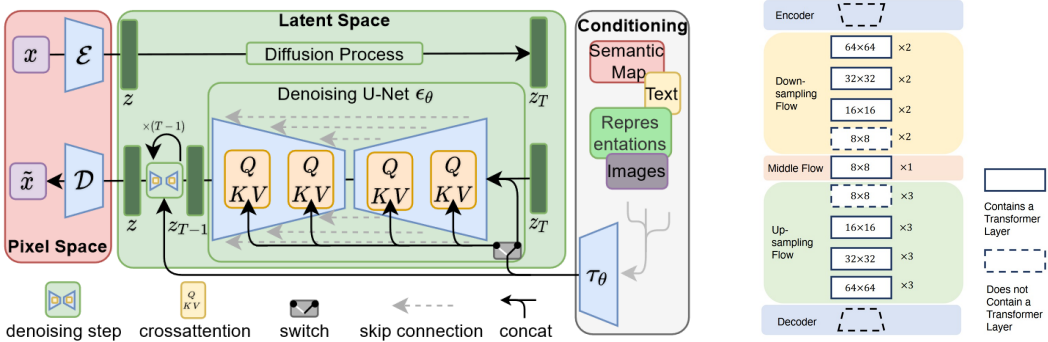


Figure: Left: LDM architecture (Rombach et al., 2022). Right: Stable Diffusion U-Net architecture (Tian et al., 2024).

Rombach, et al. High-Resolution Image Synthesis With Latent Diffusion Models, CVPR, 2022.
Tian, et al. Diffuse Attend and Segment: Unsupervised Zero-Shot Segmentation Using Stable Diffusion, CVPR, 2024.

Controllable Image Generation

Steer generation with spatial, structural, or semantic controls. ControlNet (Zhang et al., 2023) injects auxiliary conditions (e.g., edges, pose) into a frozen diffusion U-Net.

Figure: Left: Canny-edge and human-pose control. Right: ControlNet architecture (Zhang et al., 2023).

Zhang, et al. Adding Conditional Control to Text-to-Image Diffusion Models, ICCV, 2023.

Image Editing

Move from text-to-image $p(x \mid c_T)$ to instruction-guided editing $p(x \mid c_T, c_I)$, where c_I is the input image. InstructPix2Pix (Brooks et al., 2023) concatenates the input image latent with the noisy latent along the channel dimension, introducing $<$0.1M trainable parameters.

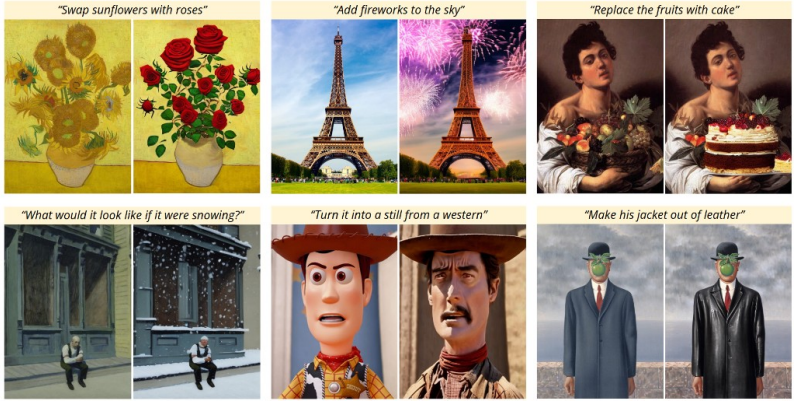


Figure: Instruction-based image editing (Brooks et al., 2023).

Brooks, et al. InstructPix2Pix: Learning To Follow Image Editing Instructions, CVPR, 2023.

Image-to-Image Translation

Translate images across domains by learning a diffusion bridge between source and target distributions.
DDBM (Zhou et al., 2024) generalizes score-based diffusion to Schrödinger-bridge image translation.

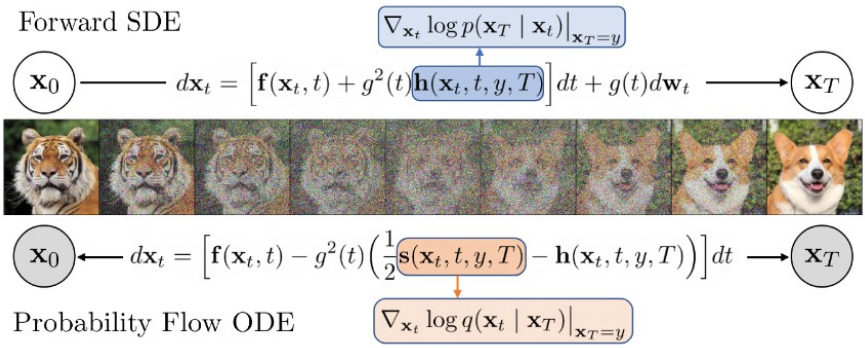


Figure: Denoising diffusion bridge from source to target image (Zhou et al., 2024).

Zhou, et al. Denoising Diffusion Bridge Models, ICLR, 2024.

Video Generation

Extend image generation to coherent temporal synthesis with efficient diffusion transformers. SANA-Video (Chen et al., 2026) uses a deep compression VAE for spatiotemporal latent modeling, together with block linear DiT and autoregressive block training for efficient text-to-video generation.

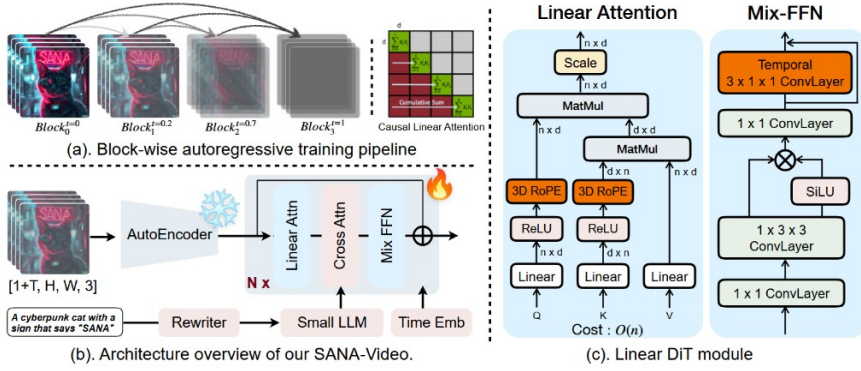


Figure: Deep compression VAE plus block linear diffusion transformer for efficient video generation (Chen et al., 2026).

Advanced Applications

Beyond standard synthesis: inversion, representation learning, and world modeling.

Inversion Task

Solve inverse problems with a pre-trained diffusion generative model—no further training required. DDRM (Kawar et al., 2022) restores images from degraded observations via diffusion priors.

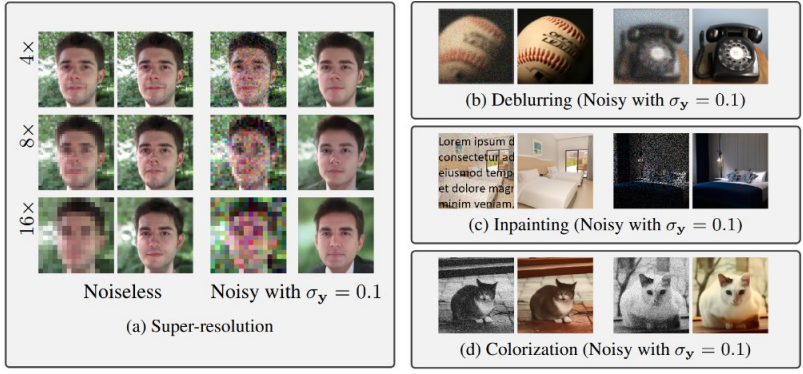


Figure: Diffusion-based restoration for super-resolution, deblurring, inpainting, and colorization (Kawar et al., 2022).

Kawar, et al. Denoising Diffusion Restoration Models, NeurIPS, 2022.

Representation Extractor

Use pre-trained diffusion models as representation extractors. DiffSeg (Tian et al., 2024) performs zero-shot segmentation from Stable Diffusion attention maps.

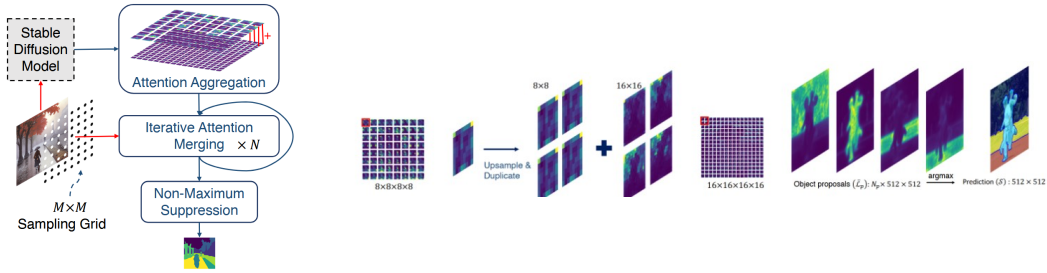


Figure: Left: DiffSeg pipeline (Tian et al., 2024). Right: (a) Attention aggregation: lower-resolution attention maps are upsampled and duplicated to match higher-resolution receptive fields. (b) NMS: maximum activation across the L_p proposals for each pixel.

Tian, et al. Diffuse Attend and Segment: Unsupervised Zero-Shot Segmentation Using Stable Diffusion, CVPR, 2024.

Generalist Vision Learner

Build unified vision systems from large-scale generative pre-training. Vision Banana (Gabeur et al., 2026) shows image generators can be generalist vision learners for both generation and understanding.

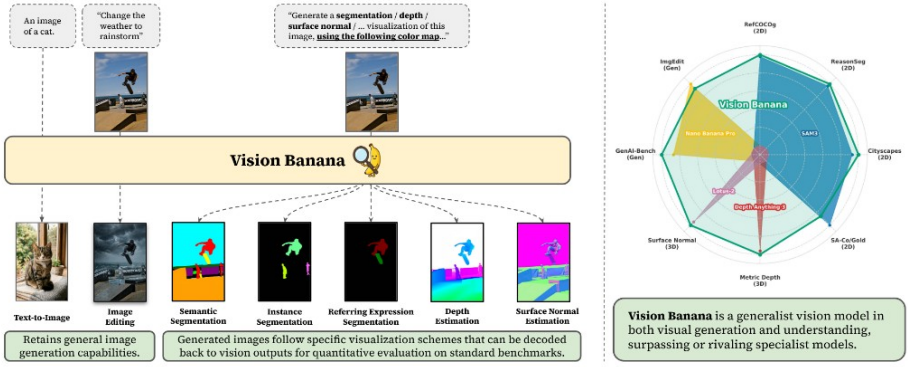


Figure: A generalist vision model for generation and perception (Gabeur et al., 2026).

Gabeur, et al. Image Generators Are Generalist Vision Learners. arXiv preprint arXiv:2604.20329. 2026.

3D/4D Generation & World Modeling

Generate navigable spatiotemporal worlds and learn interactive environment models from video. YUME 1.5 (Mao, Li, et al., 2025) extends YUME (Mao, Lin, et al., 2025) with text-controlled exploration, trained on Sekai (Li et al., 2025).

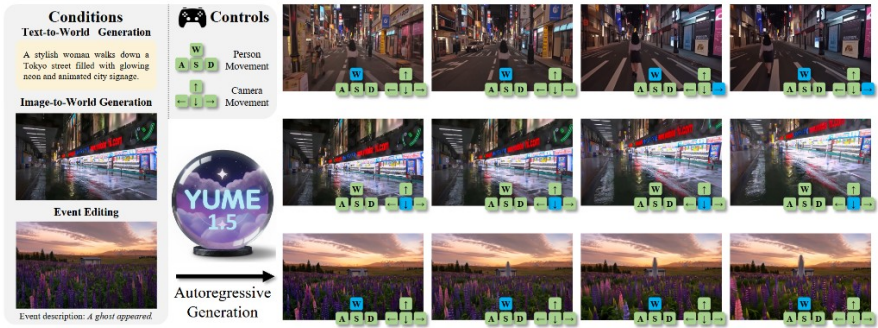


Figure: Text-controlled interactive world generation (Mao, Li, et al., 2025).

Mao, Li, et al. Yume-1.5: A Text-Controlled Interactive World Generation Model. arXiv preprint arXiv:2512.22096. 2025.
Mao, Lin, et al. Yume: An Interactive World Generation Model. arXiv preprint arXiv:2507.17744. 2025.
Li, et al. Sekai: A Video Dataset towards World Exploration, NeurIPS D&B, 2025.

Takeaways and What's Next

1. We start from an unconditional image generator $p(x)$ and progressively add more conditions to it. Technically, this is representation alignment: new condition signals—text, segmentation maps, source images, etc.—are injected into a pre-trained denoiser.
2. Besides their powerful generative capability, deep generative models are also representation learners (Yang and Wang, 2023), often matching or even surpassing SSL methods such as MoCo on downstream recognition.
3. Joint generative modeling $p(x, y)$ can natively unify generation and understanding via Bayes' rule: marginalize to obtain $p(x) = \int p(x, y)\, dy$ for synthesis, and derive $p(y \mid x) = p(x, y)/p(x)$ for discriminative inference—without training a separate head. A more robust foundation model built on this principle remains under-explored.
4. We are still on the way toward a generalist, task-agnostic, any-to-any foundation model (Gabeur et al., 2026; Zuo et al., 2025).
5. Building strong generative models is promising and already feasible, but open challenges remain in efficient training, sampling, alignment, strong generalization, and more.
6. Today's visual generation can exceed human perceptual limits—how do we properly evaluate it?

Yang, et al. Diffusion Model as Representation Learner, ICCV, 2023.
Gabeur, et al. Image Generators Are Generalist Vision Learners. arXiv preprint arXiv:2604.20329. 2026.
Zuo, et al. Is Nano Banana Pro a Low-Level Vision All-Rounder? A Comprehensive Evaluation on 14 Tasks and 40 Datasets. arXiv preprint arXiv:2512.15110. 2025.