Diffusion Models - Implementation
04/08/2026

The function $\epsilon_\theta(\mathbf{x}_t, t)$ takes as input a noisy image of the same spatial dimensions as the data, along with the timestep t, and outputs a noise estimate of identical spatial dimensions. This input-output structure — same dimensionality in and out — is naturally suited to the *U-Net* architecture (*Ronneberger et al*., 2015), originally developed for biomedical image segmentation.

U-Net Structure

The U-Net consists of:

1. *Encoder (downsampling path)*: A sequence of convolutional blocks progressively halve spatial resolution while increasing feature channel depth: $H \times W \times C \to \frac{H}{2} \times \frac{W}{2} \times 2C \to \ldots$

2. *Bottleneck*: A compact representation at the lowest spatial resolution.

3. *Decoder (upsampling path)*: Progressively restores spatial resolution via transposed convolutions or bilinear upsampling.

4. *Skip connections*: Feature maps from each encoder level are concatenated to the corresponding decoder level, preserving fine-grained spatial detail that would otherwise be lost during downsampling.

The timestep *t* is typically encoded via a sinusoidal positional embedding (analogous to transformer positional encoding):

$$\mathrm{PE}(t, 2i) = \sin\left(\frac{t}{10000^{2i/d}}\right), \quad \mathrm{PE}(t, 2i+1) = \cos\left(\frac{t}{10000^{2i/d}}\right)$$

where d is the embedding dimension. These embeddings are projected via learned linear layers and added to the feature maps at each resolution level, conditioning the network's behavior on the noise level.

Modern diffusion U-Nets (*Dhariwal & Nichol*, 2021) incorporate:

- Group normalisation and SiLU (Sigmoid-weighted Linear Unit) activations;

- Multi-head self-attention layers at lower spatial resolutions, enabling the model to capture global image structure;

- Cross-attention with text embeddings for text-conditioned generation.

Accelerated Sampling

Standard DDPM sampling requires $T = 1000$ sequential network evaluations — prohibitively slow for real-time applications. Generating a 512×512 image with $T = 1000$ and a large U-Net may take tens of seconds on modern GPUs.

© All rights reserved, Orange County Science Institute, 2026

Denoising Diffusion Implicit Models (DDIM, *Song et al.*, 2020) reformulate the reverse process as a non-Markovian deterministic update, allowing sampling in $S \ll T$ steps. The DDIM update rule is:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{predicted } \mathbf{x}_0} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \sigma_t \boldsymbol{\epsilon}_t$$

Setting $\sigma_t = 0$ makes the process fully deterministic: the same noise initialization always produces the same image. With $\sigma_t > 0$ set to recover the DDPM variance, one recovers stochastic sampling. DDIM enables 50-step sampling (20× speedup) with only modest quality degradation. The ODE underlying DDIM can be solved with higher-order numerical integrators. The DPM-Solver (*Lu et al.*, 2022) applies semi-linear ODE theory to achieve high-quality samples in 10–20 function evaluations, enabling near-real-time generation.

Conditional Generation

To condition generation on a label y (*e.g.*, ImageNet class), *Dhariwal & Nichol* (2021) proposed classifier guidance: training a separate classifier $p_\phi(y \mid \mathbf{x}_t)$ on noisy images and using its gradient to steer the reverse process:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \sqrt{1-\bar{\alpha}_t}\, \gamma\, \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}_t)$$

The scalar γ is the guidance scale, controlling the trade-off between sample fidelity (adherence to the condition y) and sample diversity.

Classifier-free guidance (*Ho & Salimans*, 2022) eliminates the need for a separate classifier by jointly training a conditional model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)$ and an unconditional model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset)$ (obtained by randomly dropping the conditioning signal c during training). At inference:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, c) = (1+\gamma)\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c) - \gamma\, \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \emptyset)$$

This is equivalent to moving the predicted noise estimate away from the unconditional direction and towards the conditional direction, amplifying adherence to the condition c (*e.g.*, a text prompt). Classifier-free guidance is the mechanism responsible for the remarkable prompt-following behavior of systems like DALL·E 2 and Stable Diffusion.

In text-to-image systems, the conditioning signal c is a text embedding produced by a pretrained CLIP encoder (Contrastive Language–Image Pretraining, *Radford et al.*, 2021). CLIP is trained on image–text pairs to align their representations in a shared embedding space via contrastive learning:

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{e^{\text{sim}(\mathbf{z}_i^I, \mathbf{z}_i^T)/\tau}}{\sum_j e^{\text{sim}(\mathbf{z}_i^I, \mathbf{z}_j^T)/\tau}} + \log \frac{e^{\text{sim}(\mathbf{z}_i^T, \mathbf{z}_i^I)/\tau}}{\sum_j e^{\text{sim}(\mathbf{z}_j^I, \mathbf{z}_i^T)/\tau}} \right]$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a learnable temperature, and $(\mathbf{z}^I, \mathbf{z}^T)$are image and text embeddings respectively. The text embedding $\mathbf{z}^T$ is injected into the U-Net via cross-attention layers.

© All rights reserved, Orange County Science Institute, 2026

Latent Diffusion Models

Applying the diffusion process directly in pixel space for high-resolution images (***e.g.***, $512 \times 512 \times 3$) is computationally prohibitive: the U-Net must process $\sim 786{,}000$-dimensional vectors, and the self-attention layers exhibit $\mathcal{O}(n^2)$ complexity in spatial resolution.
Latent Diffusion Models (LDMs, *Rombach et al.*, 2022) — the architecture underlying Stable Diffusion — address this by performing the diffusion process in the compressed latent space of a pretrained Variational Autoencoder (VAE).

The VAE consists of:

- An encoder $\mathcal{E}$: $\mathbf{x} \mapsto \mathbf{z} = \mathcal{E}(\mathbf{x})$, compressing $512 \times 512 \times 3$ images to $64 \times 64 \times 4$ latent codes (a factor-$48\times$ reduction in dimensionality);

- A decoder $\mathcal{D}$: $\mathbf{z} \mapsto \tilde{\mathbf{x}} = \mathcal{D}(\mathbf{z})$, reconstructing pixel-space images from latents.

The diffusion model is trained entirely on latent codes $\mathbf{z}$:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(\mathbf{x}_0),\, \boldsymbol{\epsilon},\, t}\Big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)\|^2\Big]$$

At sampling time, the reverse process generates a latent $\hat{\mathbf{z}}_0$, which is decoded to pixel space by $\mathcal{D}$. This reduces computational cost by orders of magnitude while preserving perceptual quality — the VAE compresses away imperceptible high-frequency detail while retaining semantically meaningful structure.

Evaluation Metrics

Evaluating generative models involves measuring two competing objectives: fidelity (do samples look realistic?) and diversity (do samples cover the full data distribution?).

The Fréchet Inception Distance (*Heusel et al.*, 2017) is the dominant quantitative metric. It compares the distributions of real and generated images in the feature space of a pretrained Inception-v3 network, modeling each as a multivariate *Gaussian*:

$$\text{FID} = \left\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\right\|^2 + \text{Tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{1/2}\Big)$$

where $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ and $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ are the mean and covariance of real and generated feature distributions respectively. Lower FID indicates greater similarity between distributions. State-of-the-art diffusion models achieve FID < 2 on standard benchmarks (***e.g.***, CIFAR-10, ImageNet 256×256), compared to ~ 10–30 for early GAN models.

The Inception Score (*Salimans et al.*, 2016) measures both sharpness (the classifier should be confident about generated image content) and diversity (the marginal class distribution should be uniform):

$$\text{IS} = \exp(\mathbb{E}_{\mathbf{x}}[D_{\text{KL}}(p(y \mid \mathbf{x}) \,\|\, p(y))])$$

© All rights reserved, Orange County Science Institute, 2026

Higher IS indicates better quality. IS does not capture mode dropping as effectively as FID and has fallen out of favor as the primary benchmark.

Fundamental Assumptions: A Critical Appraisal

The diffusion model framework rests on the following assumptions, each of which warrants critical examination:

	Assumption	Justification	Limitation
1	Gaussian forward process	Tractable marginals; closed-form reverse posteriors	May not optimally model non-Gaussian data (*e.g.*, discrete tokens)
2	Small step size ($\beta_t \ll 1$)	Guarantees Gaussian approximation of reverse steps	Requires large T; slow sampling
3	Markov structure	Simplifies likelihood decomposition	Restricts dependencies across timesteps
4	Gaussian reverse approximation	Empirically validated; theoretically justified for small β_t	Approximation quality degrades at large β_t
5	Universal function approximation	Neural networks can approximate score functions	Requires large models and massive data
6	Closed-world data distribution	Training data covers the target distribution	Out-of-distribution generalization is limited

© All rights reserved, Orange County Science Institute, 2026