CVPR 2022

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1], Andreas Blattmann[1], Dominik Lorenz[1], Patrick Esser[2], Björn Ommer[1]

[1] LMU Munich & IWR, Heidelberg University [2] Runway ML

Paper Code

Problem

Diffusion models set the state of the art in image synthesis, but they run **directly in pixel space** — so training the most powerful models costs **hundreds of GPU-days**, and every sampling step is slow and memory-hungry.

This confines high-resolution generative modeling to the few labs with massive compute, leaving a large carbon footprint and limiting access for the broader research community. Sampling is also **sequential**, so each image needs many costly denoising passes.

> *Training a top pixel-space diffusion model can take 150–1000 V100-days — out of reach for all but a handful of labs.*

Motivation

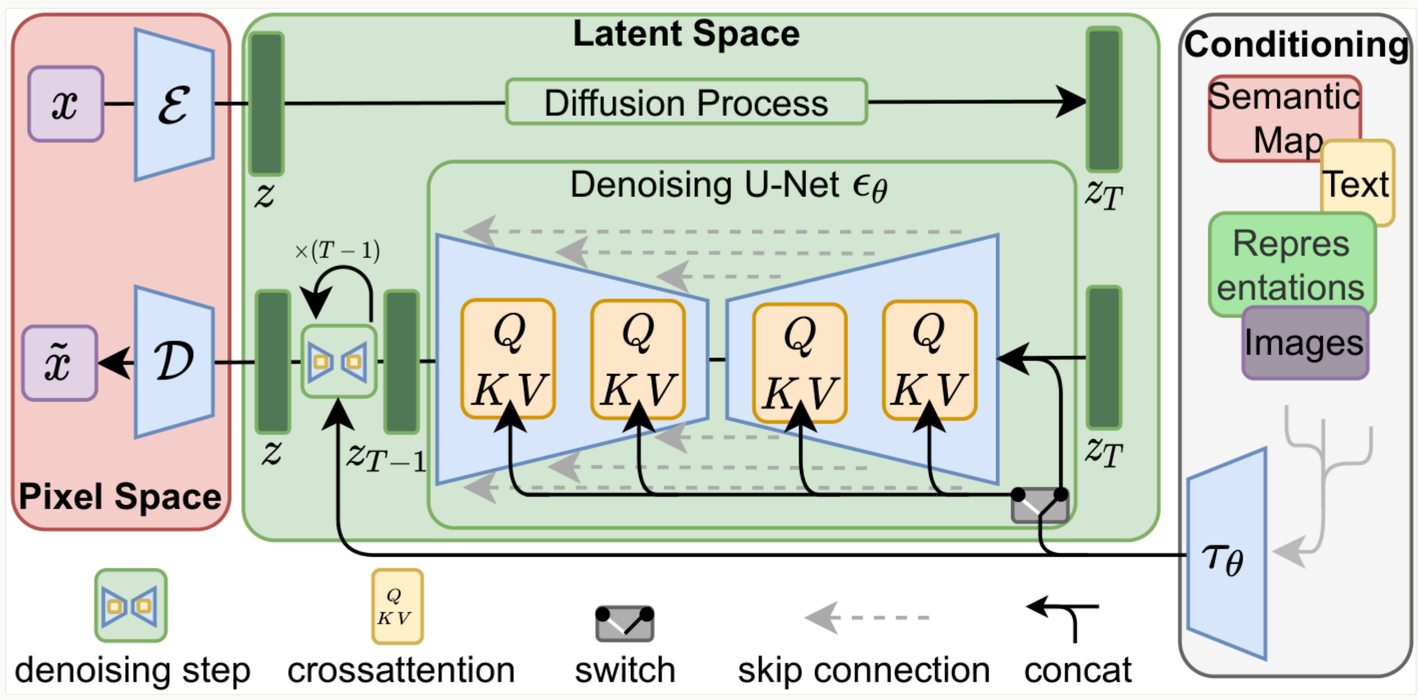
- Most bits of an image encode **imperceptible high-frequency detail**; a likelihood model spends the bulk of its capacity and compute on pixels, not semantics.
- Prior diffusion work down-weighted those details in the loss, yet the network still had to be evaluated on **every pixel** — so the wasted computation remained.

PIXEL SPACE
Spends capacity on imperceptible detail; evaluates every pixel.

vs.

LATENT SPACE
Train where semantics live — a compact, perceptual code.

METHOD



LDM architecture — an autoencoder maps pixels ↔ latent; a UNet denoises in latent space, conditioned via cross-attention (τθ).

- **Stage 1 — autoencoder.** A perceptual autoencoder (KL- or VQ-regularized, perceptual + patch-adversarial loss) compresses images by a factor **f** into a compact latent.
- **Stage 2 — latent diffusion.** A time-conditional UNet runs the diffusion process entirely in that latent space, where every forward pass is far cheaper than in pixels.
- **Conditioning.** Cross-attention layers inject text, layouts, or class labels, turning the UNet into a general-purpose multi-modal generator.

1 Encode → latent › **2** Diffuse in latent › **3** Cross-attn condition

Headline Numbers

5.11

FID · CelebA-HQ 256² (LDM-4)

New state of the art on CelebA-HQ.

3.60	1×A100	f = 4–8	≥1.6×
FID · ImageNet (LDM-4-G)	all main experiments	best compression	inpainting FID gain

Key Results

Class-cond. ImageNet	FID ↓	Params
ADM	10.94	554M
LDM-4 (ours, no cfg)	10.56	400M
ADM-G	4.59	608M
LDM-4-G (ours)	**3.60**	**400M**

−0.99 FID vs ADM-G — with 34% fewer parameters

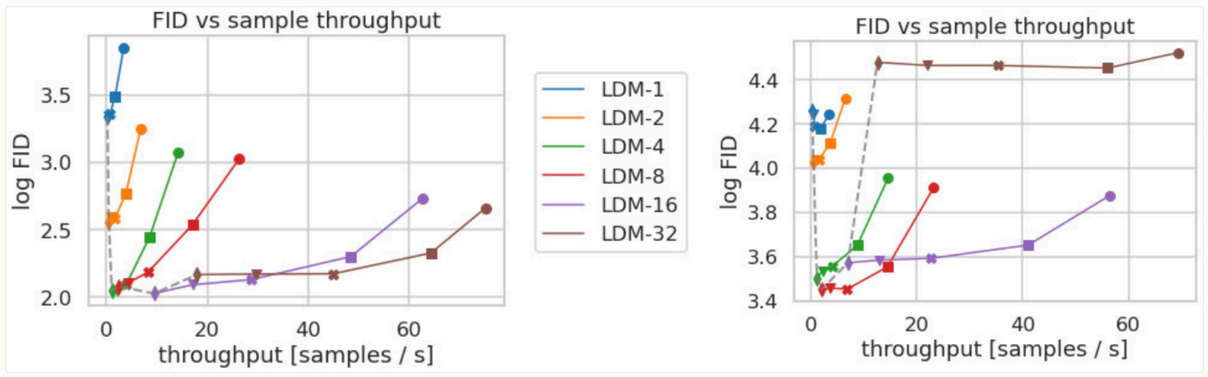
SO WHAT → *LDM-4-G reaches near-SOTA class-conditional ImageNet FID using ~⅔ the parameters and a fraction of pixel-space compute.*

Ablation Study

- **Compression sweet spot:** f = 4 and f = 8 reach far lower FID and faster sampling than pixel-space LDM-1; aggressive f = 16/32 discard too much detail.
- **Data-dependent:** complex ImageNet benefits from more compression than simple face data; VQ vs. KL latents trade reconstruction against sample FID.

The autoencoder's compression factor — not the diffusion model — is the decisive lever on quality and speed.

SO WHAT → *Mild perceptual compression (f = 4–8) is the key lever balancing efficiency and quality — neither pixel-space LDM-1 nor aggressive f = 32 competes.*



FID vs. sampling throughput across compression factors on CelebA-HQ (left) and ImageNet (right) — mild compression (LDM-4/8) gives the best quality–speed trade-off.

Also SOTA in inpainting, beats SR3 in super-resolution FID, and powers text- & layout-to-image synthesis.