CVPR 2022 POSTER

High-Resolution Image Synthesis with Latent Diffusion Models

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

[1] Ludwig Maximilian University of Munich & IWR, Heidelberg University [2] Runway ML

Paper Code

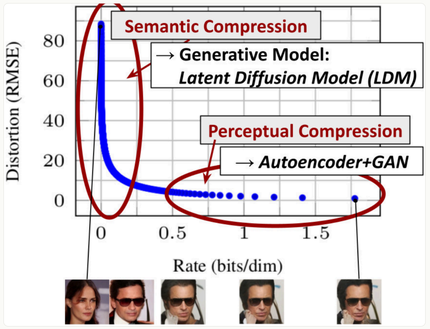
Problem

Diffusion models lead image synthesis, but run **directly in pixel space** — training the most powerful models costs **150–1000 GPU days**, and sampling is slow and memory-heavy.

Most of that compute models detail humans cannot even perceive.

Motivation

- Most image bits are **imperceptible high-frequency detail**; a likelihood model wastes its capacity there, not on semantics.
- Prior work down-weighted that, but the net still **ran on every pixel** — wasted compute remained.



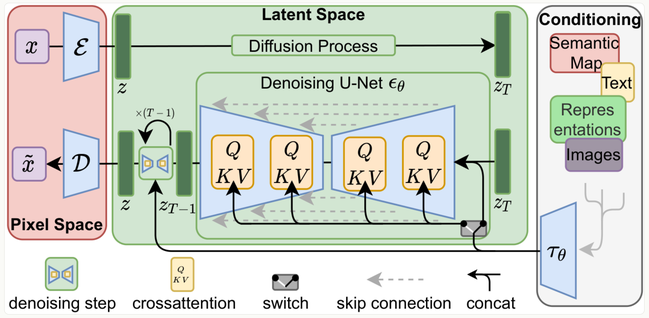
Most image bits are imperceptible perceptual detail; semantic structure is where generative capacity belongs.

Method

1 Encode → z › 2 Latent diffusion › 3 Decode → x̃

- **Stage 1 — Perceptual autoencoder.** Compress images by a factor **f** into a latent using a perceptual + patch-adversarial loss, regularized by a slight KL penalty or vector quantization.
- **Stage 2 — Latent diffusion.** Train a time-conditional UNet entirely in that compact latent space, where each denoising step is far cheaper than in pixel space.
- **Cross-attention conditioning** injects text, layouts, or class labels at every UNet layer; fully-convolutional sampling lifts 256² training to megapixel synthesis.

Mild downsampling (f=4–8) hits the sweet spot between complexity reduction and detail preservation.



LDM architecture: an autoencoder maps pixels ↔ latent z; a UNet runs diffusion in latent space, conditioned via cross-attention on text, layouts, images, or class labels.

Dataset / Benchmark

Evaluated across generation, super-resolution, and inpainting — on a single A100 GPU.

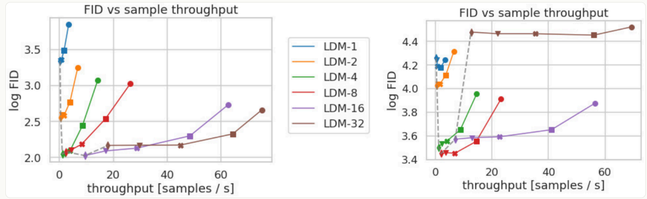
CelebA-HQ · FFHQ · LSUN-Churches · LSUN-Bedrooms · ImageNet · COCO · Places · FID · IS · P/R · PSNR/SSIM/LPIPS

Key Results

SETTING	FID ↓
CelebA-HQ · unconditional	**5.11**
ImageNet · class-cond. (CFG)	3.60
ImageNet · LDM-8-G variant	7.76

New SOTA FID 5.11 on CelebA-HQ · 3.60 on class-conditional ImageNet

SO WHAT → *LDMs match or beat pixel-space diffusion and GANs at a fraction of the training and inference cost.*



FID vs. throughput by compression factor (CelebA-HQ left, ImageNet right): f=4–8 dominates; f=1 and f=16/32 lag.

Also SOTA on inpainting; super-resolution beats SR3 in FID.

Ablation Study

DOWNSAMPLING F	OUTCOME
f=1 (pixel) / f=2	slow, worse FID
f=4 – f=8	**best FID + fast sampling**
f=16 / f=32	detail lost, capped

Headline Numbers

5.11

FID · CelebA-HQ (new SOTA)

beats prior GANs & likelihood models

3.60 FID · ImageNet (CFG)

1×A100 single-GPU training

f=4–8 best compression

Takeaway

Run diffusion in a compact, perceptually-faithful latent space — not raw pixels — cutting training and inference cost at equal-or-better quality; cross-attention adds general conditioning.

LEGACY From GPU-days to one GPU — the basis for Stable Diffusion.