DDPM (U-Net, noise ($\varepsilon$), etc.)

↳ DDIM (non-stoch) → improved DDPM  fewer noise timesteps, etc.

Latent (LDM): pixel space → latent space (w/ percept. compression and encoder), Q, K, V made using input vectors, trained weights, added back (res-block)

into. cross-attention $\left(Att. = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V\right)$

encode noisy img → time → transformer block → unet → $\hat{\varepsilon}$ latent — trained independently

text input → (tokenized, encoded)

( tr. used on low-res steps

Stable Diffusion: using LDM, CLIP

↳ SDXL: larger model, additional conditioning parameters (h/w) w/ t-emb (cropping)

better training data (more sizes), better text encoder

transformer: add position encoding to img patches, send through transformer blocks, condition on adaLN, to-patch

uses latent.

video: 16 frames / training data, similar mechanics as DDPM U-Net,