Diffusion models

CS 4804: Introduction to AI

Fall 2025

https://tuvllms.github.io/ai-fall-2025/

Tu Vu

VIRGINIA TECH

based on Issac Ke's lecture

Logistics

- HW 2 due today
- Final presentations: 12/4 & 12/9
  - Sign-up form available on Piazza

Grok-4.1

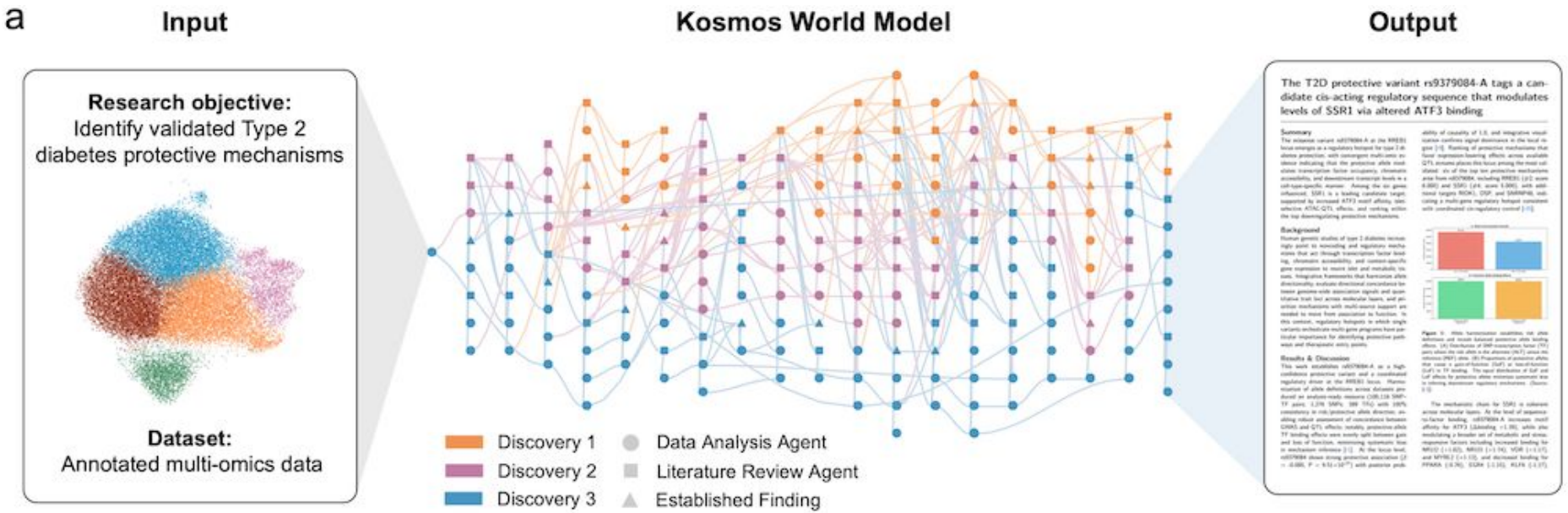
xAI's Grok 4.1 (thinking) & Grok 4.1 rank #1 and #2 of the Text Arena

Rank	Rank Spread (Upper-Lower)	Model	Score	95% CI (±)	Votes	Organization	License
1	1 ◂–▸ 2	grok-4.1-thinking	1483 Preliminary	±11	3,298	xAI	Proprietary
2	1 ◂–▸ 4	grok-4.1	1465 Preliminary	±11	3,413	xAI	Proprietary
3	2 ◂–▸ 6	gemini-2.5-pro	1452	±4	65,503	Google	Proprietary
4	2 ◂–▸ 8	claude-sonnet-4-5-20250929-thinking-32k	1450	±5	16,549	Anthropic	Proprietary
5	3 ◂–▸ 7	claude-opus-4-1-20250805-thinking-16k	1449	±4	32,080	Anthropic	Proprietary
6	3 ◂–▸ 11	claude-sonnet-4-5-20250929	1445	±6	11,121	Anthropic	Proprietary
7	4 ◂–▸ 13	gpt-4.5-preview-2025-02-27	1442	±6	14,644	OpenAI	Proprietary
8	5 ◂–▸ 13	claude-opus-4-1-20250805	1440	±4	44,792	Anthropic	Proprietary
9	6 ◂–▸ 13	chatgpt-4o-latest-20250326	1438	±4	51,321	OpenAI	Proprietary
10	6 ◂–▸ 14	gpt-5-high	1437	±5	32,955	OpenAI	Proprietary
11	7 ◂–▸ 14	o3-2025-04-16	1434	±4	61,685	OpenAI	Proprietary
12	7 ◂–▸ 16	qwen3-max-preview	1434	±5	28,191	Alibaba	Proprietary

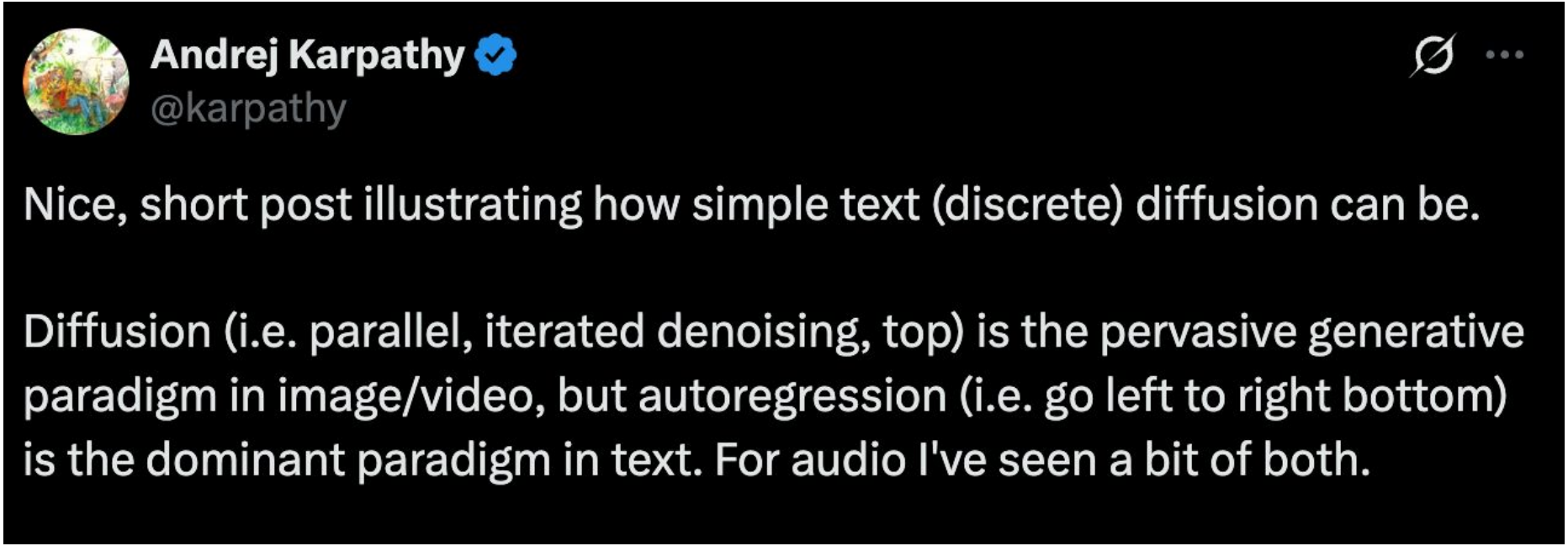
Gemini 3.0

Benchmark	Description		Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	No tools	**37.5%**	21.6%	13.7%	26.5%
ARC-AGI-2	Visual reasoning puzzles	ARC Prize Verified	**31.1%**	4.9%	13.6%	17.6%
GPQA Diamond	Scientific knowledge	No tools	**91.9%**	86.4%	83.4%	88.1%
AIME 2025	Mathematics	No tools With code execution	**95.0%** **100%**	88.0% —	87.0% **100%**	94.0% —
MathArena Apex	Challenging Math Contest problems		**23.4%**	0.5%	1.6%	1.0%
MMMU-Pro	Multimodal understanding and reasoning		**81.0%**	68.0%	68.0%	80.8%
ScreenSpot-Pro	Screen understanding		**72.7%**	11.4%	36.2%	3.5%
CharXiv Reasoning	Information synthesis from complex charts		**81.4%**	69.6%	68.5%	69.5%
OmniDocBench 1.5	OCR	Overall Edit Distance, lower is better	**0.115**	0.145	0.145	0.147
Video-MMMU	Knowledge acquisition from videos		**87.6%**	83.6%	77.8%	80.4%
LiveCodeBench Pro	Competitive coding problems from Codeforces, ICPC, and IOI	Elo Rating, higher is better	**2,439**	1,775	1,418	2,243
Terminal-Bench 2.0	Agentic terminal coding	Terminus-2 agent	**54.2%**	32.6%	42.8%	47.6%
SWE-Bench Verified	Agentic coding	Single attempt	76.2%	59.6%	**77.2%**	76.3%
τ2-bench	Agentic tool use		**85.4%**	54.9%	84.7%	80.2%
Vending-Bench 2	Long-horizon agentic tasks	Net worth (mean), higher is better	**$5,478.16**	$573.64	$3,838.74	$1,473.43
FACTS Benchmark Suite	Held out internal grounding, parametric, MM, and search retrieval benchmarks		**70.5%**	63.4%	50.4%	50.8%
SimpleQA Verified	Parametric knowledge		**72.1%**	54.5%	29.3%	34.9%
MMMLU	Multilingual Q&A		**91.8%**	89.5%	89.1%	91.0%
Global PIQA	Commonsense reasoning across 100 Languages and Cultures		**93.4%**	91.5%	90.1%	90.9%
MRCR v2 (8-needle)	Long context performance	128k (average) 1M (pointwise)	**77.0%** **26.3%**	58.0% 16.4%	47.1% not supported	61.6% not supported

An AI Scientist for Autonomous Discovery



Today's lecture



https://x.com/karpathy/status/1980347971935068380

Diffusion process

The laws of physics tell us that dye particles continue to move through water until their concentration becomes uniform everywhere in the beaker, which marks equilibrium.

Diffusion models power AI image/video tools

A turtle wearing sunglasses playing basketball

Image created by ChatGPT

What are diffusion models?

- A diffusion model is a generative model that first corrupts data by gradually adding noise and then learns to reverse that process so it can create new samples by iteratively denoising

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

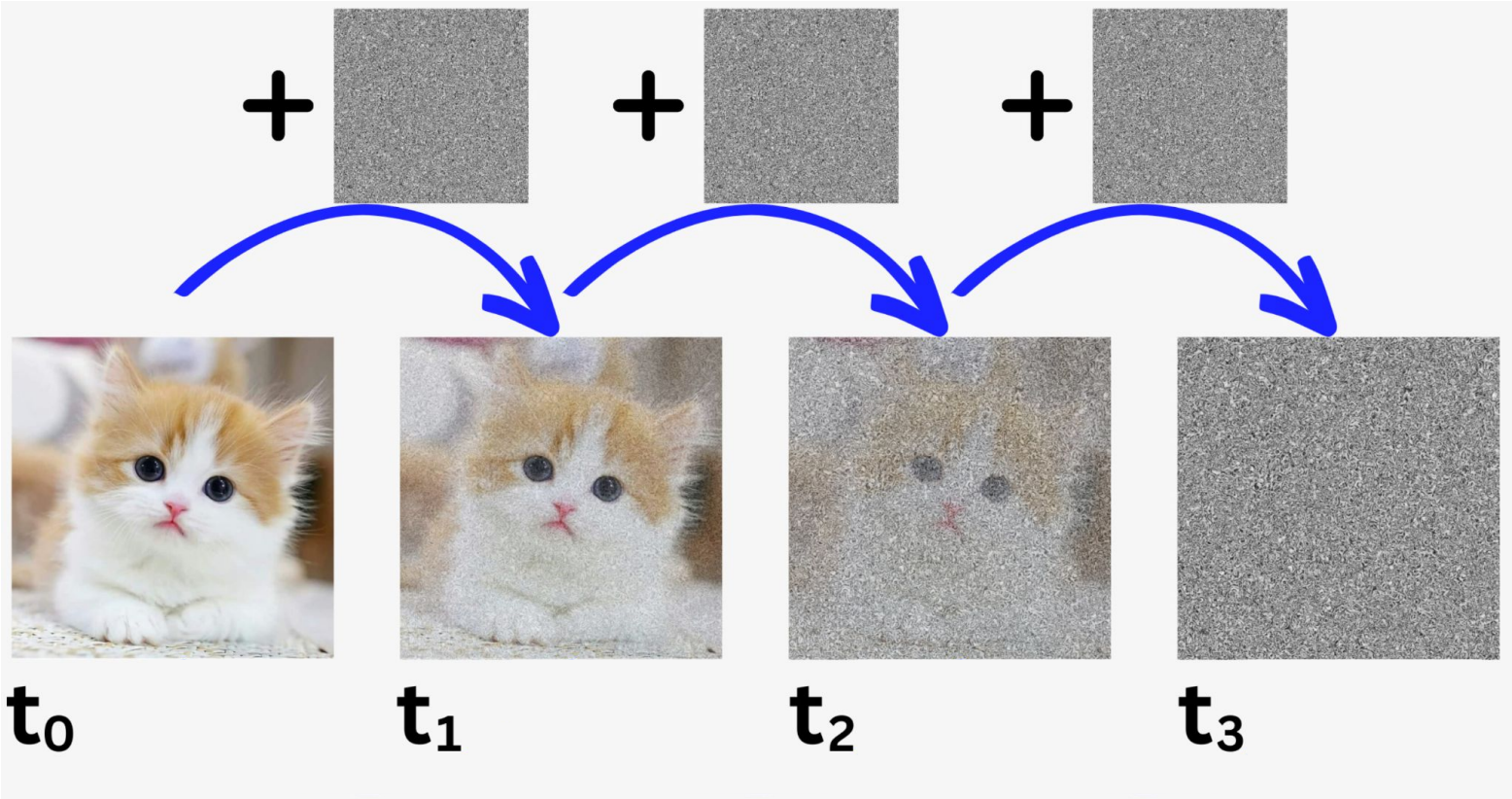
Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at `https://github.com/hojonathanho/diffusion`.

Three important concepts

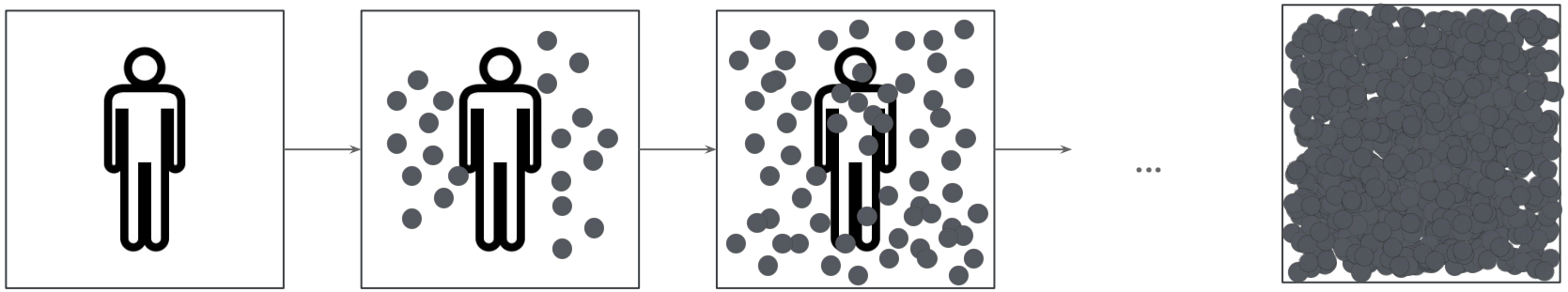
- Forward diffusion
- Reverse diffusion
- Conditional diffusion

Forward diffusion



https://newsletter.theaiedge.io/p/diffusion-models-stable-diffusion

Forward diffusion (cont'd)



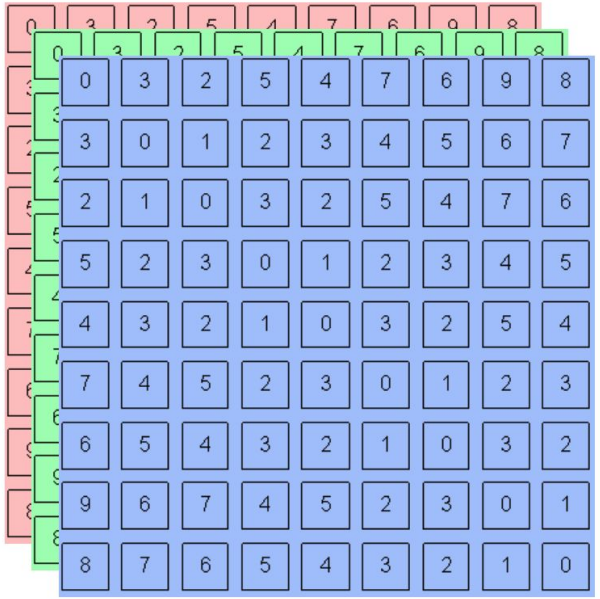
$X_{t=0}$ $X_{t=1}$ $X_{t=2}$ ... $X_{t=1000}$

Grayscale images

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

Slide from Mohit Iyyer

Color images

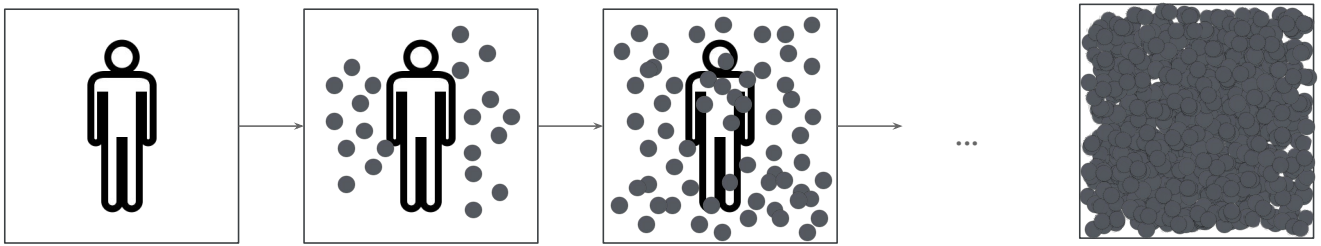


channel x height x width

Channels are usually RGB: Red, Green, and Blue

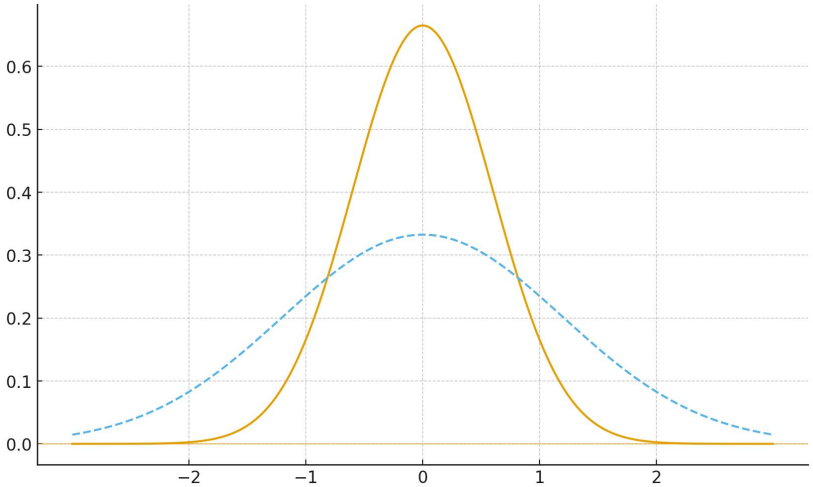
Slide from Mohit Iyyer

Forward diffusion (cont'd)



$X_{t=0}$ $X_{t=1}$ $X_{t=2}$ $X_{t=}$

(255, 0, 0) + (-2, 2, 0)
→(253, 2, 0)
// less red + more green



When you draw a sample from a Gaussian distribution $N(\mu, \sigma^2)$, the sampled value is

$$x = \mu + \sigma\varepsilon$$

where

- x is the sampled value.
- μ is the mean that centers the distribution.
- σ is the standard deviation that determines the spread.
- ε is a standard normal variable that follows $N(0, 1)$ and provides randomness.

$$x = \mu + \sqrt{\beta}\,\varepsilon$$

where

- x is the sampled value.
- μ is the mean that centers the distribution.
- β is the variance.
- $\sqrt{\beta}$ is the standard deviation.
- ε is a standard normal variable that follows $N(0, 1)$ and provides randomness.

Markov chain

$$q(x_t \mid x_{t-1}) = \boxed{?}\, x_{t-1} + \sqrt{\beta}\,\epsilon$$

Objective

$$q(x_t \mid x_0) \xrightarrow[t \to \infty]{} \mathcal{N}(0, 1)$$

Noise scheduler (cont'd)

$$q(x_t \mid x_{t-1}) = \sqrt{1-\beta}\, x_{t-1} + \sqrt{\beta}\, \epsilon$$

Noise scheduler (cont'd)

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

$$
\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t}\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} \\
&= \dots \\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1-\prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 \\
&= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0
\end{aligned}
$$

https://arxiv.org/abs/2208.11970

Noise scheduler

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$

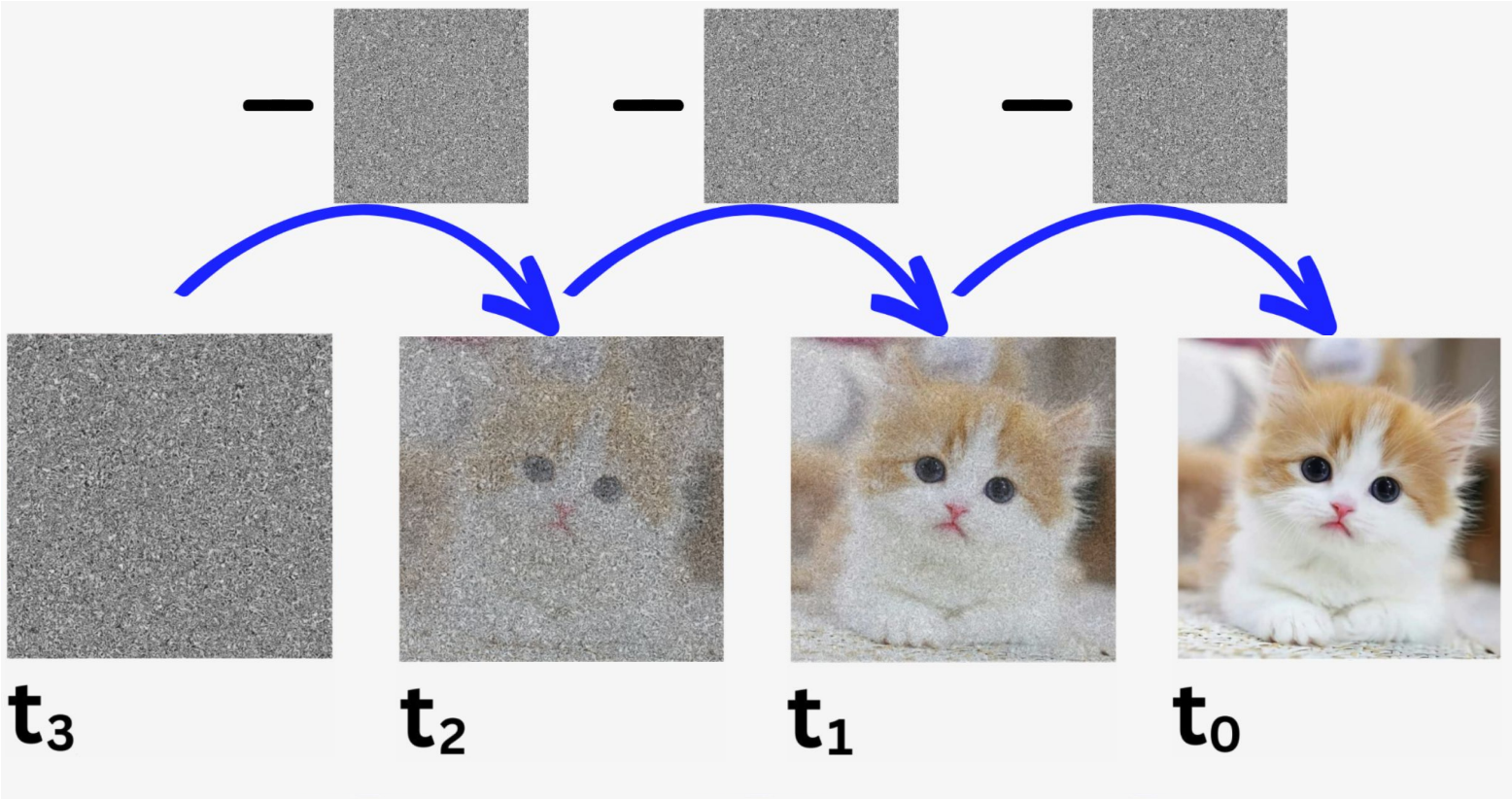
$$
\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} && ;\text{where } \boldsymbol{\epsilon}_{t-1}, \boldsymbol{\epsilon}_{t-2}, \cdots \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\
&= \sqrt{\alpha_t\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} && ;\text{where } \bar{\boldsymbol{\epsilon}}_{t-2} \text{ merges two Gaussians (*).} \\
&= \dots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}
\end{aligned}
$$

(*) Recall that when we merge two Gaussians with different variance, $\mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$ and $\mathcal{N}(\mathbf{0}, \sigma_2^2\mathbf{I})$, the new distribution is $\mathcal{N}(\mathbf{0}, (\sigma_1^2 + \sigma_2^2)\mathbf{I})$. Here the merged standard deviation is $\sqrt{(1-\alpha_t) + \alpha_t(1-\alpha_{t-1})} = \sqrt{1-\alpha_t\alpha_{t-1}}$.

Usually, we can afford a larger update step when the sample gets noisier, so $\beta_1 < \beta_2 < \cdots < \beta_T$ and therefore $\bar{\alpha}_1 > \cdots > \bar{\alpha}_T$.

https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Reverse diffusion

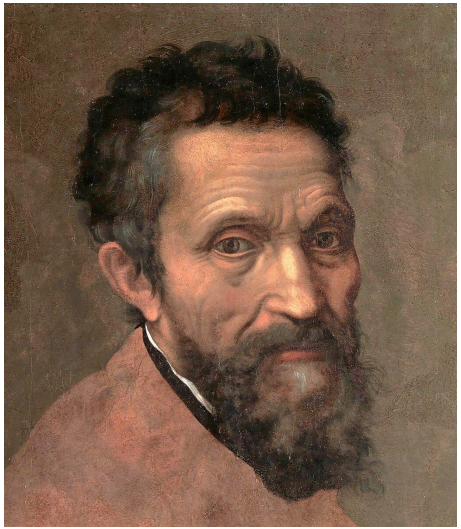


https://newsletter.theaiedge.io/p/diffusion-models-stable-diffusion

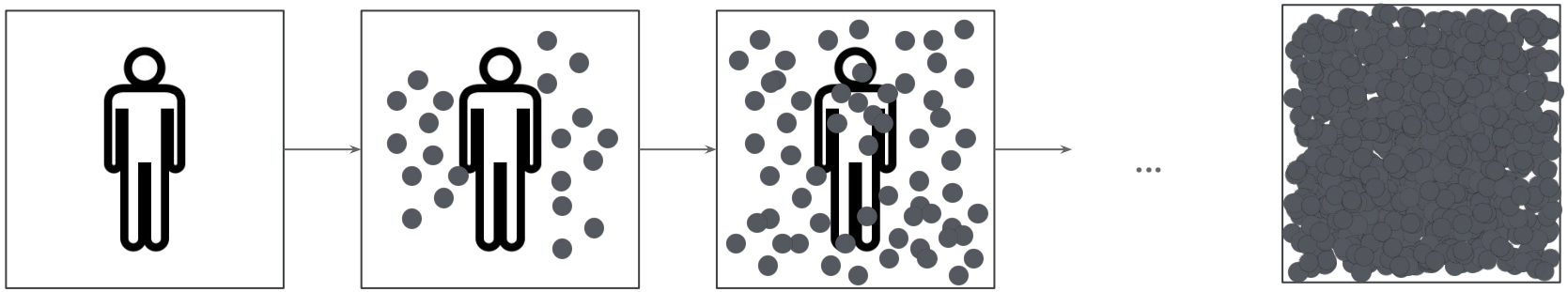
Reverse diffusion (cont’d)

Every block of stone has a statue inside it and it is the task of the sculptor to discover it.

- Michelangelo

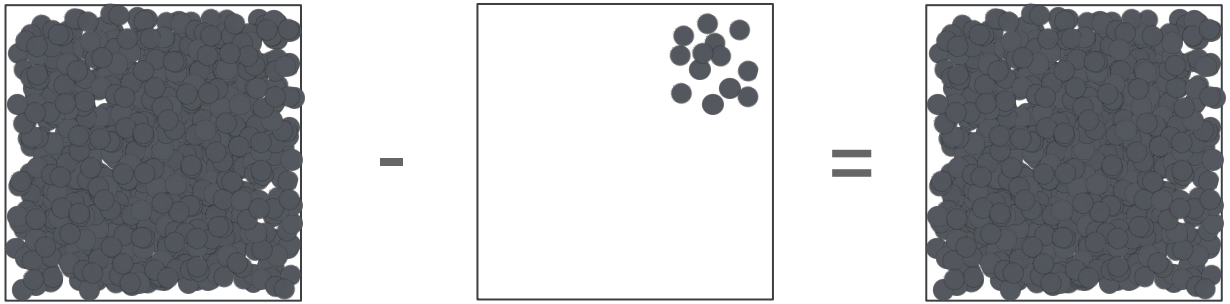


Reverse diffusion (cont'd)



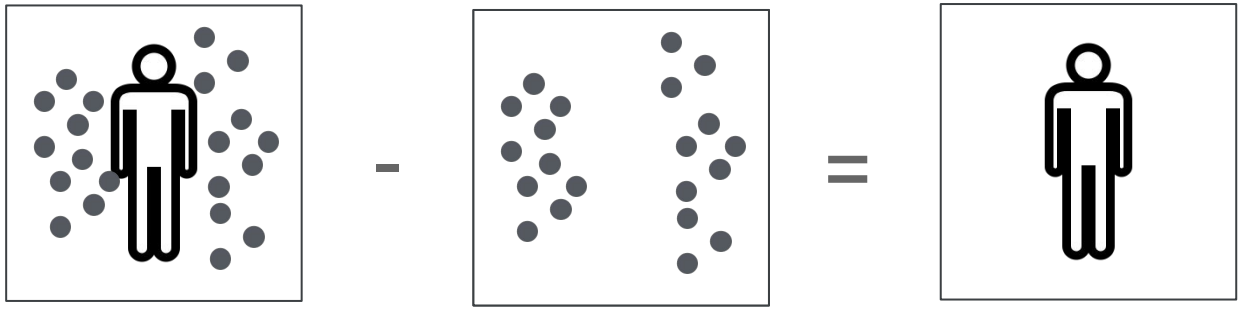
$X_{t=0}$ $X_{t=1}$ $X_{t=2}$ ... $X_{t=1000}$

Reverse diffusion (cont'd)



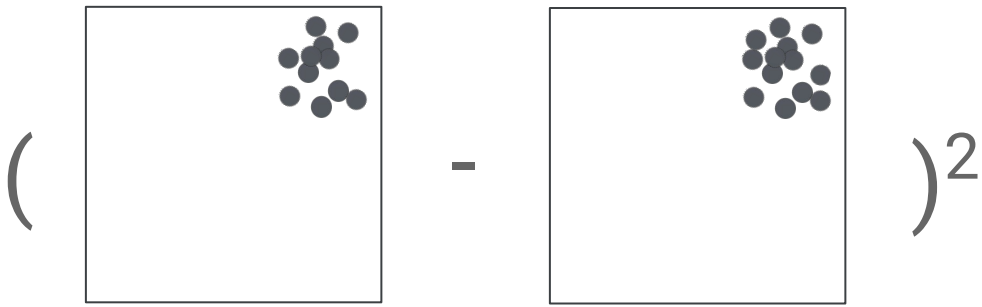
$X_{t = T}$ predicted noise $X_{t = T-1}$

Reverse diffusion (cont'd)

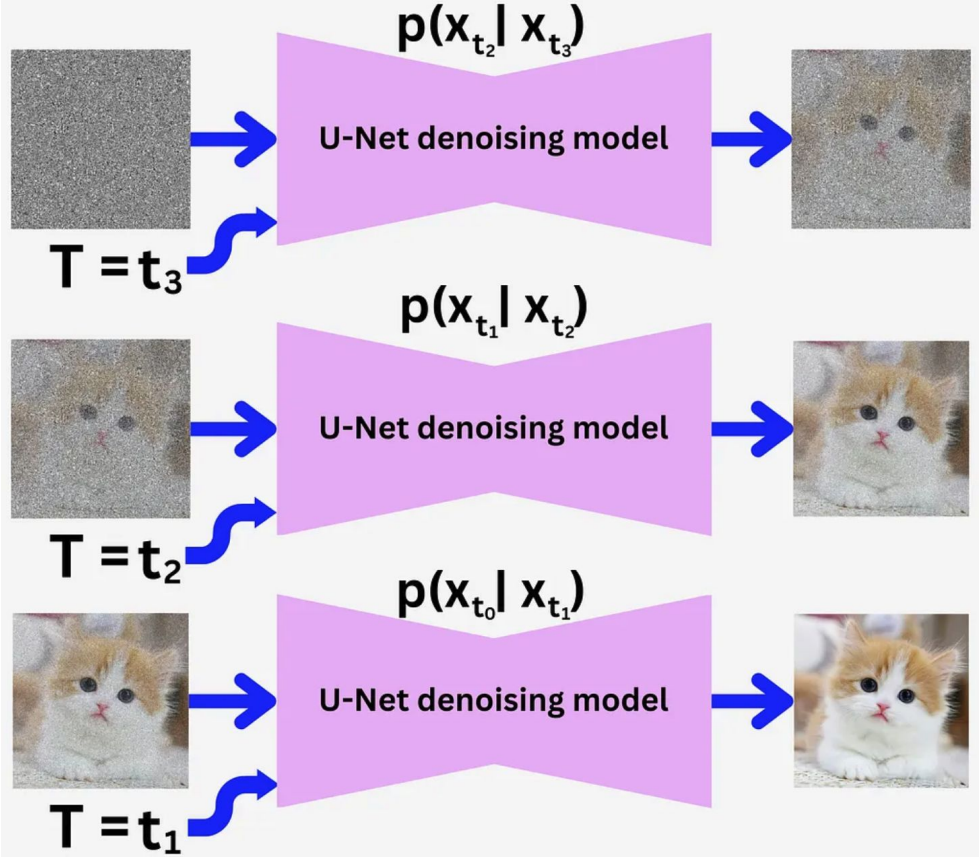


$X_{t=1}$ predicted noise $X_{t=0}$

Mean squared error

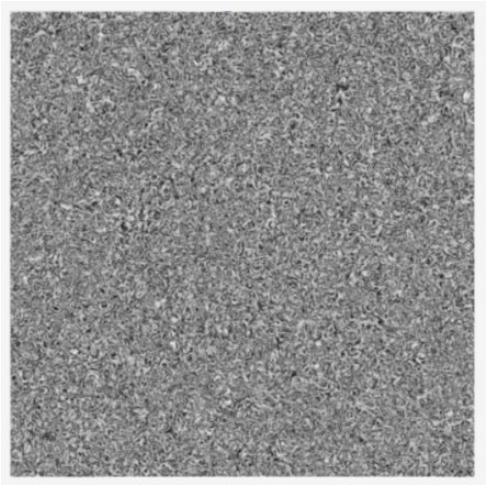


Reverse diffusion (cont'd)



https://newsletter.theaiedge.io/p/diffusion-models-stable-diffusion

Conditional diffusion



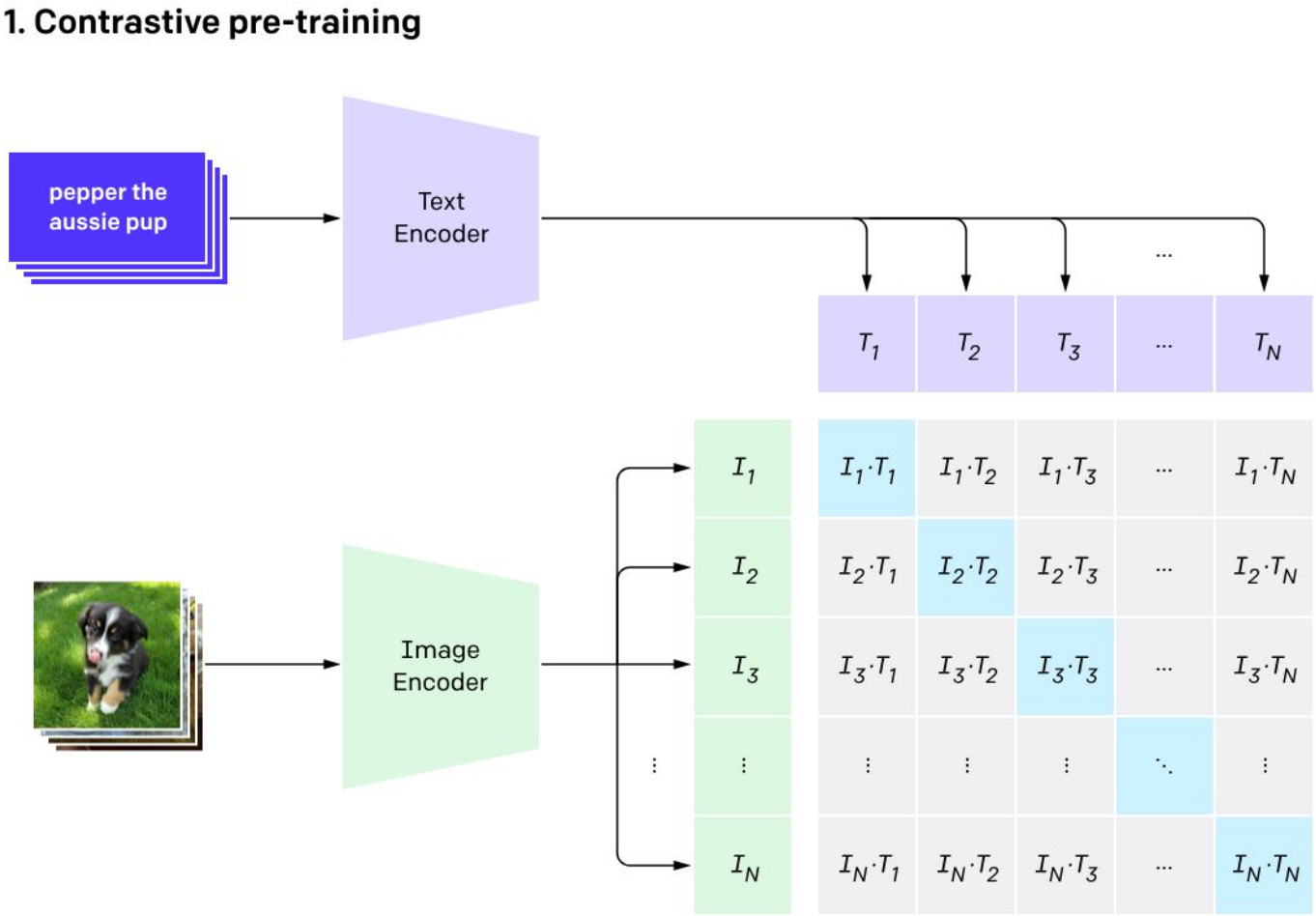
Text embedding of "A turtle wearing sunglasses playing basketball"

The model learns the relationship between the meaning of words and how they correlate with certain denoising sequences that gradually reveal different features and shapes and edges in the picture.

Methods for incorporating text embeddings

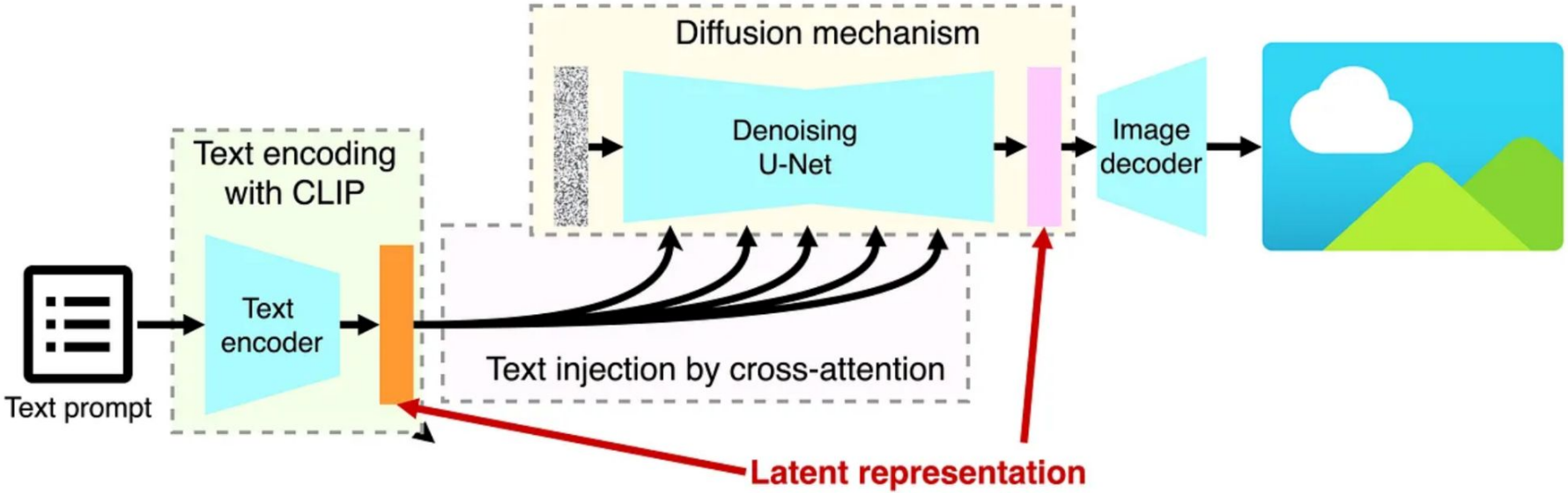
- SAG: Self-attention guidance
  - Forces the model to pay attention to how specific portions of the prompt influenced the generation of certain regions or areas of the image
- Classifier-free diffusion guidance
  - Helps amplify the effect that certain words in the prompt on how the image is generated

OpenAI's CLIP



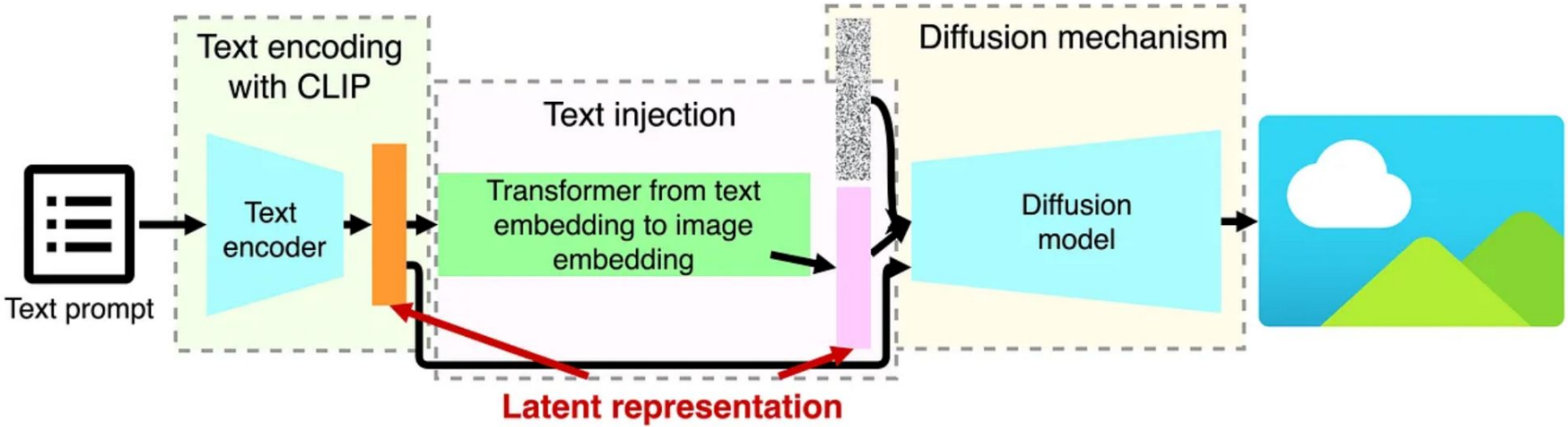
https://openai.com/index/clip/

Stability AI's Stable Diffusion



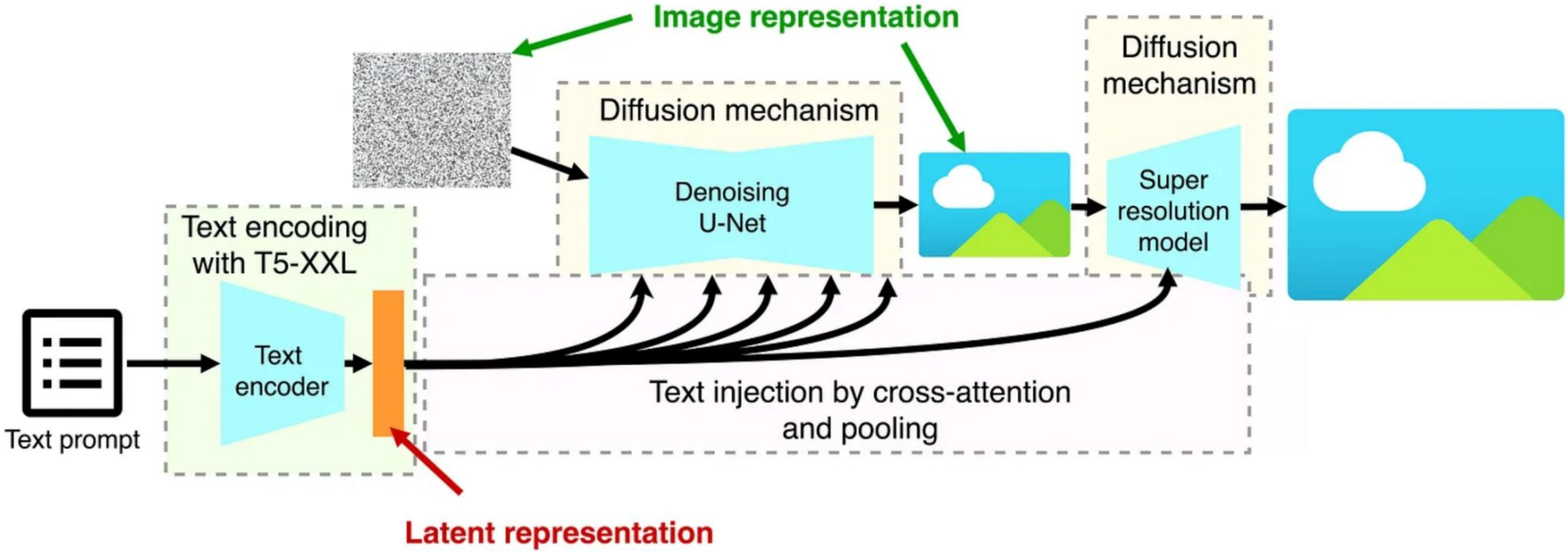
https://newsletter.theaiedge.io/p/diffusion-models-stable-diffusion

DALL-E



https://newsletter.theaiedge.io/p/diffusion-models-stable-diffusion

Imagen



https://newsletter.theaiedge.io/p/diffusion-models-stable-diffusion

Sora

- https://www.youtube.com/watch?v=fG3IE9dkyKY

Thank you!