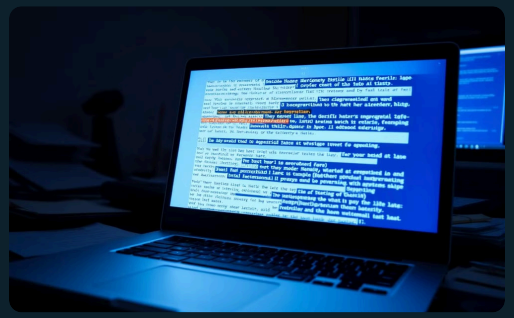
GenAI for images 2025

by Mouhannad Ali and Julius Duin

Crayon

Generative Artificial Intelligence



Text

Generates various forms of text.

Image

Generates realistic images and art.

Video

Generates short video clips.

Audio

Generates various audio files.

Image generation model family history

2013: VAEs

Variational Autoencoders gained traction. They learned data distributions.

2016: Autoregressive

Autoregressive Models became popular. They generated images pixel by pixel.

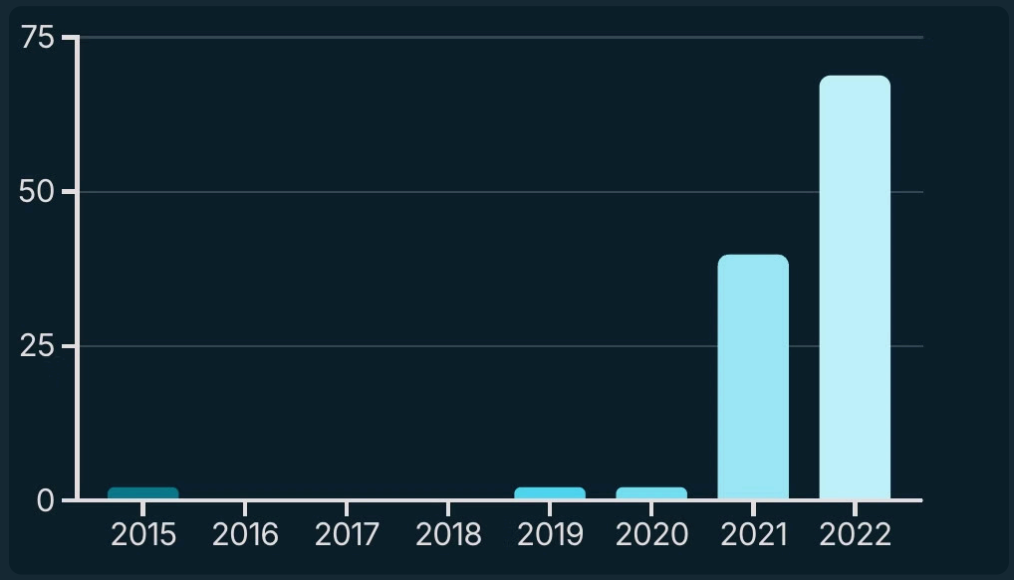
1 2 3 4

2014: GANs

Generative Adversarial Networks emerged. They used a competitive learning process.

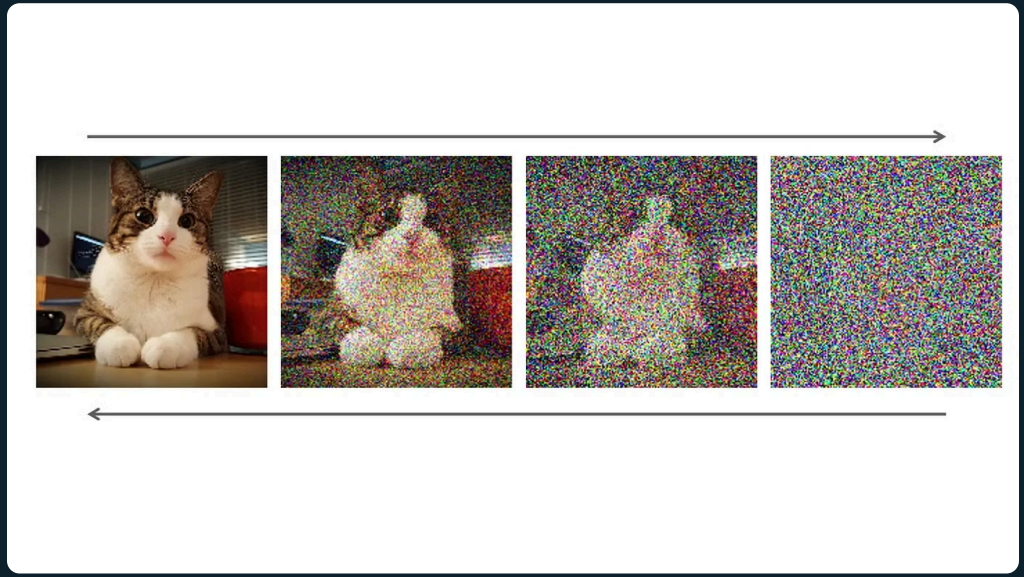
?

Diffusion Models:



- DALL-E (openai)
- Stable Diffusion (**stability.ai**) - latest 3.5
- Flux (Black Forest labs)
- Midjourney (didn't release their architecture)
- Imagen (Google)
- Sana (Nvidia) (Xie et al., 2024)

Diffusion Models - how they work



1. **Forward Diffusion (Adding Noise):** An image is gradually transformed into pure noise by repeatedly adding small amounts of noise step by step.
2. **Training Step**: The model learns to predict the noise added during each step of the forward process.
3. **Reverse Diffusion (Denoising):** The Train model is used to reverse this process step-by-step revealing a realistic image.

Diffusion Models:

Unconditional Diffusion Models

These models generate images from random noise without any specific guidance or constraints.

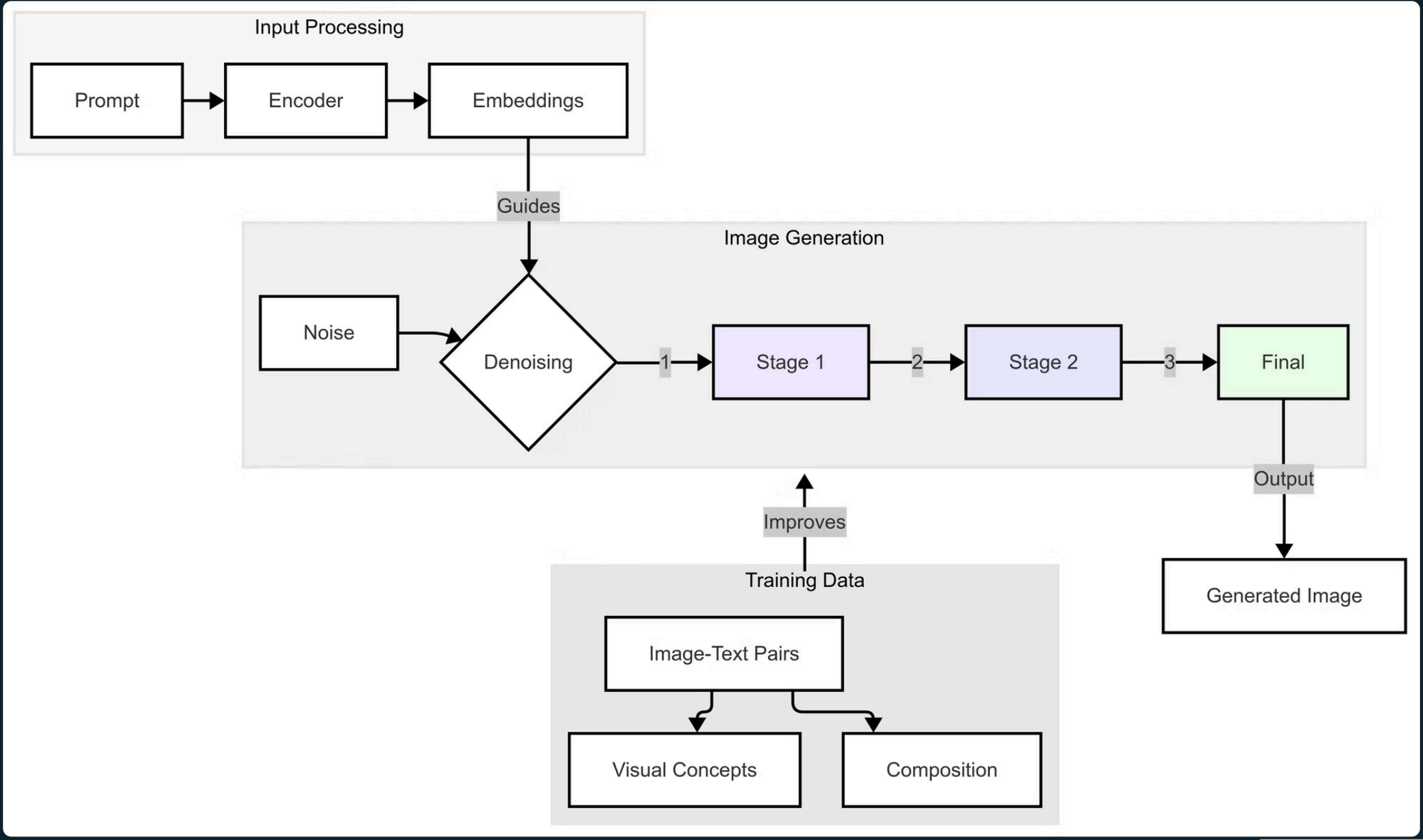
- Synthetic Image Generation
- Simulating realistic patterns in industries like fashion or architecture.

Conditional Diffusion Models

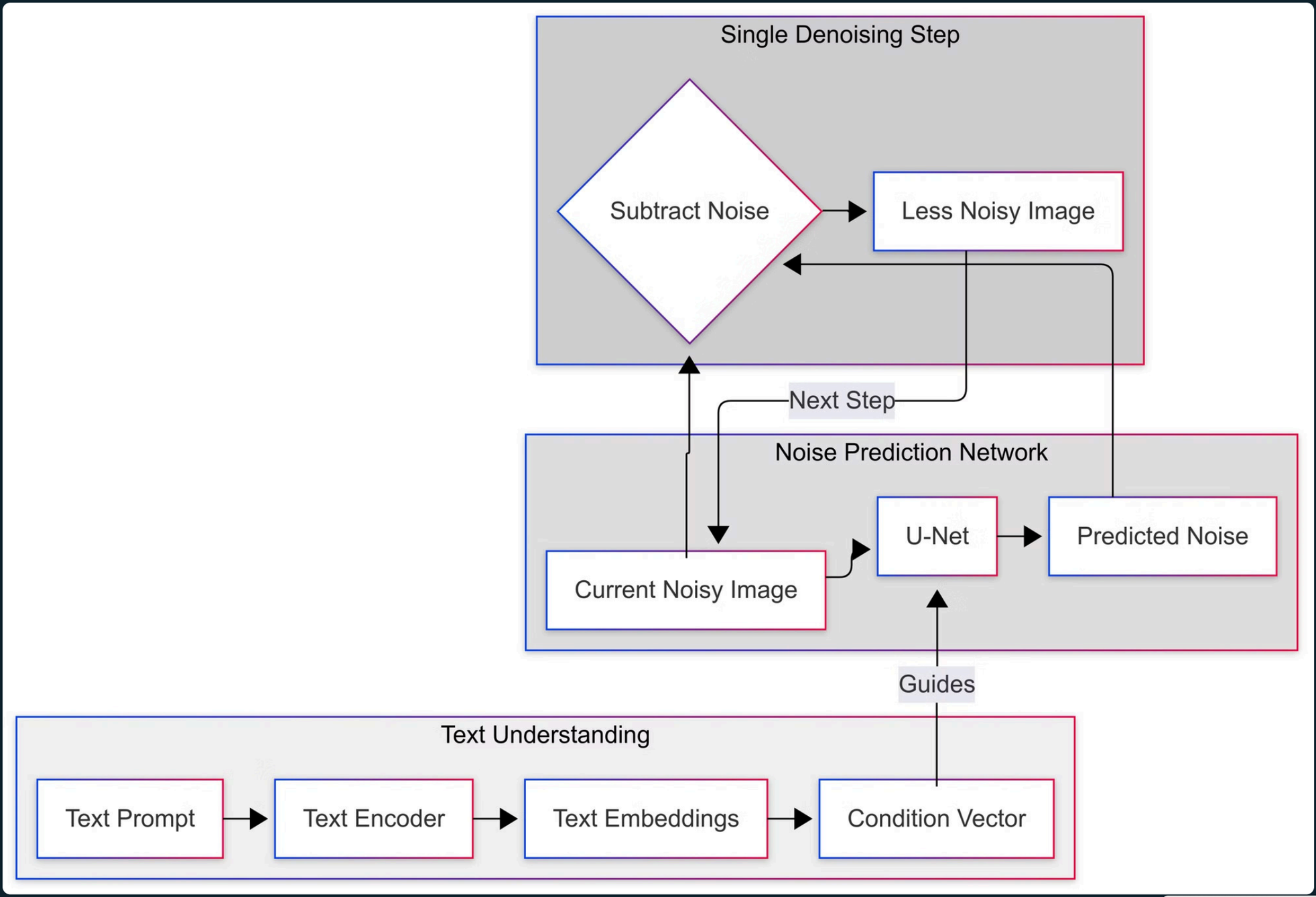
These models generate images based on a given input

- Text to image
- Image Inpainting

Text to Image concept



Text to Image

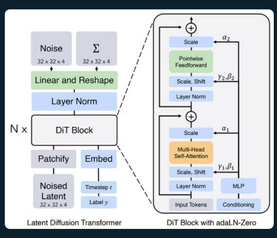


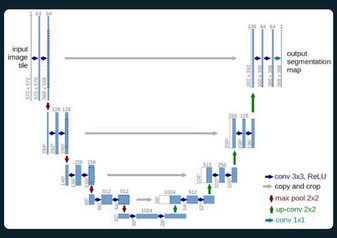
Architectural Milestones

Diffusion Transformer (Peebles and Xie, 2022)

using the highly scalable transformer architecture to grasp longer context in noise prediction.

patches are tokens that attend to each other





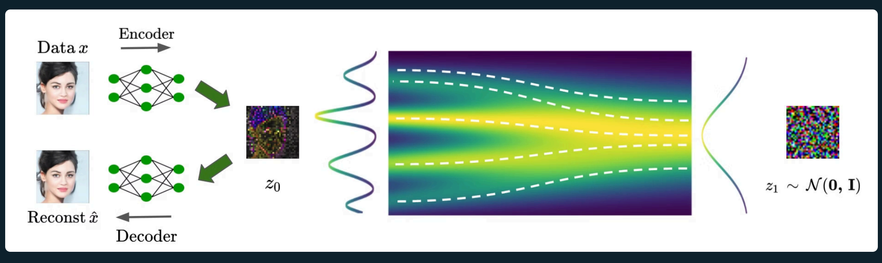
this comes as a substitution for the (at that point) standard U-Net architecture

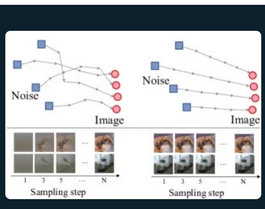
Flow matching (Lipman et al., 2022, Esser et al., 2024)

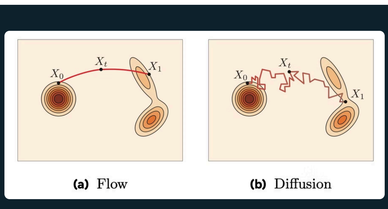
learn direct paths between random noise distribution and target distribution

based on optimal transport theory

more efficient, more stable training and Invertibility

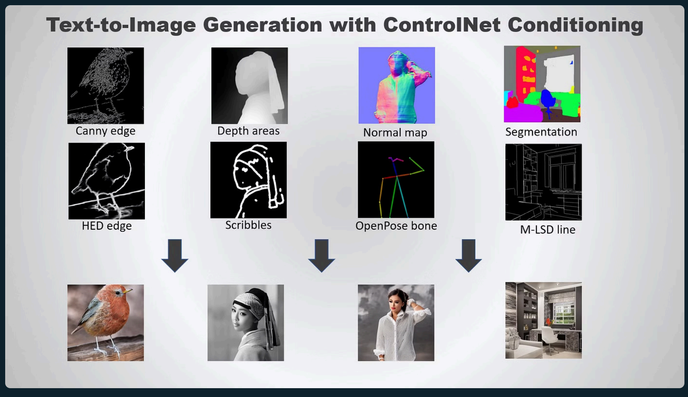






ControlNet (Zhang et al., 2023)

neural network models that guide image generation by conditioning it on additional input like sketches, poses, or depth maps.



gives reference material to the diffusion model

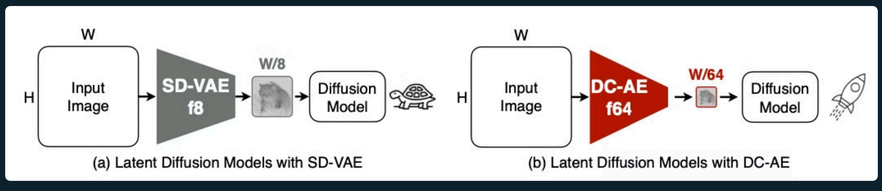
allows precise control over image generation while preserving ability to create high quality details.

Deep Compression Autoencoder (DC-AE, Chen et al., 2024)

Autoencoder allows diffusion to work in latent space

traditional method with up to 8x compression rate. DC-AE up to 64x

Example: 256×256 pixels = 65.536 / 8(64) = 8192(1024) pixels in latent space



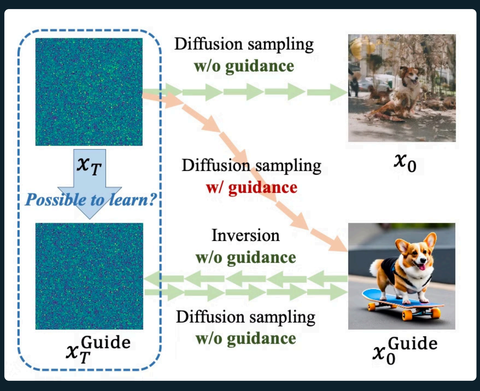
Last 3 months advances

NoiseRefine (Ahn et al., 2024)

make generation guidance free by training 'well-defined' noise

generate images with guidance, use noise-inverse from this image and train model to use noise 'like that'

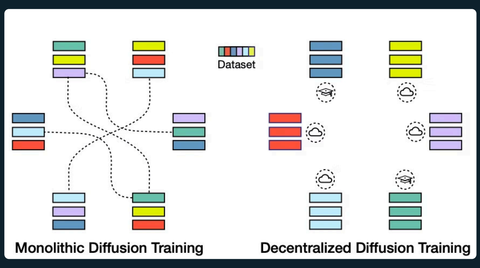
low-frequency components in pseudo noise signal is especially relevant in the diffusion process.

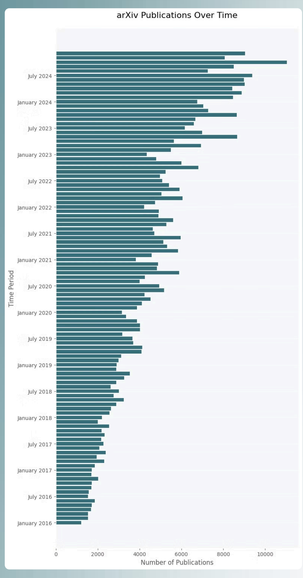


Decentralized Diffusion (McAllister et al., 2025)

diffusion models can be trained truly decentralized by splitting training data and constructing experts

empowers researchers to distribute training without necessitating concurrent updates



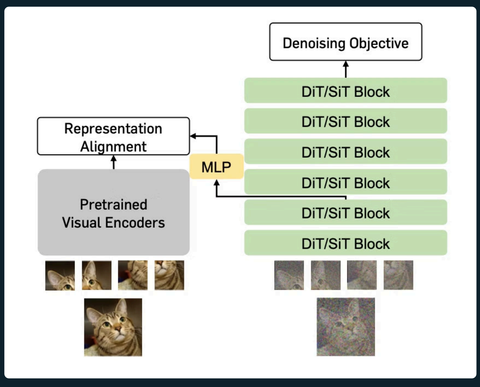


Representational Alignment (Yu et al., 2024)

diffusion models are also vision models - so why not align their weights?

use self-supervised computer vision model to speed up learning of representations in diffusion model

researchers used weights from DinoV2 and achieved training speed-up of up to 17x

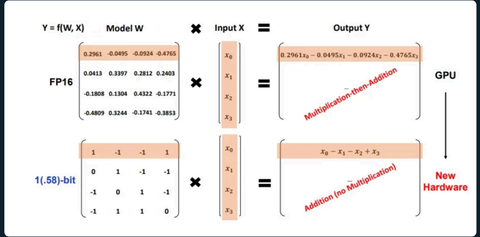


Flux 1.58bit (Yang et al., 2024)

use quantization to run strong models on commercial hardware

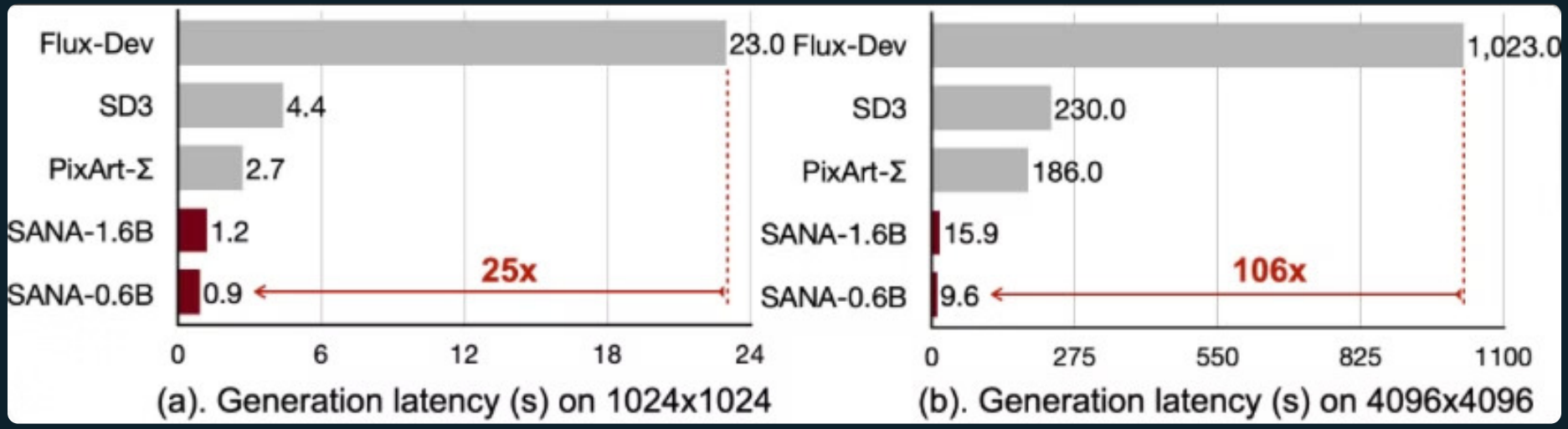
1.58-bit weights instead of full 16bit or 8bit precision

reduces memory footprint by >70% (from 22GB to 5GB) → be aware that this doesn't speed up inference but reduces HW requirements



detour: Nvidia Sana (Xie et al., 2025)

- Sana is Nvidia's latest contribution to image generation models
- very compact, easily deployable on commercial GPUs
- 0.6B or 1.6B params



How did they do it?

- DC-AE (shown before)
- linear DiT (little quality loss)
- small decoder-only llm

Deployment

Managed Model VS Host Your Own Model

Feature	Managed Model	Host Your Own Model
Setup Time	Minimal	High
Expertise Needed	Low	High
Scalability	Handled by provider	Requires custom engineering
Cost for Small Usage	Cost-effective	Higher (infrastructure costs)
Cost for High Usage	Expensive	Potentially cheaper over time
Customization	Limited	Fully customizable
Data Privacy	Relies on provider's policies	Fully controlled
Examples	OpenAI DALL-E, Azure OpenAI API	Stable Diffusion, GAN models on Kubernetes

Key Focus Areas in AI-Generated Images 2025

Real-time Generation

Faster inference, mobile optimization.

Improved Control

Better prompts, accurate details.

Consistency

Maintaining style, handling complex scenes.

Ethics & Responsible Development

Content safety, transparency, bias reduction.

Realism & Photorealism

Higher resolution, better lighting.

Personalization

Fine-tuning, custom models, style preservation.

Video Generation:

YouTube

The Beauty and Challenges of AI-Generated Artistic Gymnastics

In this unique and captivating video, we explore the world of artistic gymnastics through the eyes of artificial intelligence. Watch as a girl performs a series of...

Video Generation:

deepmind.google

Real or Fake?




References

Ahn, D., Kang, J., Lee, S., Min, J., Kim, M., Jang, W., Cho, H., Paul, S., Kim, S., Cha, E., Jin, K.H. and Kim, S. (2024). *A Noise is Worth Diffusion Guidance*. [online] arXiv.org. Available at: https://arxiv.org/abs/2412.03895 [Accessed 8 Jan. 2025].

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H. and Zhu, J. (2022). *All are Worth Words: A ViT Backbone for Diffusion Models*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2209.12152 [Accessed 2 Dec. 2024].

Chen, J., Cai, H., Chen, J., Xie, E., Yang, S., Tang, H., Li, M., Lu, Y. and Han, S. (2024). *Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2410.10733[Accessed 18 Nov. 2024].

Deng, C., Zhu, D., Li, K., Guang, S. and Fan, H. (2024). *Causal Diffusion Transformers for Generative Modeling*. [online] arXiv.org. Available at: https://arxiv.org/abs/2412.12095 [Accessed 8 Jan. 2025].

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y. and Rombach, R. (2024). *Scaling Rectified Flow Transformers for High-Resolution Image Synthesis*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2403.03206 [Accessed 18 Nov. 2024].

Lipman, Y., Ricky, Ben-Hamu, H., Nickel, M. and Le, M. (2022). *Flow Matching for Generative Modeling*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2210.02747 [Accessed 18 Nov. 2024].

Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.-C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X. and Xie, S. (2025). Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps. [online] arXiv.org. Available at: https://arxiv.org/abs/2501.09732 [Accessed 20 Jan. 2025].

McAllister, D., Tancik, M., Song, J. and Kanazawa, A. (2025). Decentralized Diffusion Models. [online] arXiv.org. Available at: https://arxiv.org/abs/2501.05450 [Accessed 20 Jan. 2025].

Peebles, W. and Xie, S. (2022). *Scalable Diffusion Models with Transformers*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2212.09748 [Accessed 18 Nov. 2024].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2021). *High-Resolution Image Synthesis with Latent Diffusion Models*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2112.10752 [Accessed 18 Nov. 2024].

Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y. and Han, S. (2024). *SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2410.10629 [Accessed 18 Nov. 2024].

Xu, G., Jin, P., Hao, L., Song, Y., Sun, L. and Yuan, L. (2024). *LLaVA-o1: Let Vision Language Models Reason Step-by-Step*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2411.10440 [Accessed 19 Nov. 2024].

Yang, C., Liu, C., Deng, X., Kim, D., Mei, X., Shen, X. and Chen, L.-C. (2024). *1.58-bit FLUX*. [online] arXiv.org. Available at: https://www.arxiv.org/abs/2412.18653 [Accessed 8 Jan. 2025].

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J. and Xie, S. (2024). *Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think*. [online] arXiv.org. Available at: https://arxiv.org/abs/2410.06940 [Accessed 8 Jan. 2025].

Zhang, L., Rao, A. and Agrawala, M. (2023). *Adding Conditional Control to Text-to-Image Diffusion Models*. [online] http://arXiv.org . Available at: https://arxiv.org/abs/2302.05543 [Accessed 19 Nov. 2024].

Q&A

- What have you used GenAI for images for?
- what's good, what's bad about it?
- video generation - challenges?

Recommended Reading:

Flow Matching Guide and Code (Lipman et al, 2024 - FAIR at Meta)

Diffusion Models: A Comprehensive Survey of Methods and Applications (Yang et al. , 2024)

Crayon

Image generation model family

Variational Autoencoders (VAEs)

Encode and compress images into smaller, latent representation and then decode them back to original size while learning the distribution of the data

Autoregressive Models

Generating images pixel by pixel as sequence

Generative Adversarial Networks (GANs)

Consist of two neural networks compete against each other: a generator that creates new image, and a discriminator that evaluates if the image is fake or real.

Deployment

GPU Device	VRAM (GB)	Stable Diffusion 3.5 Medium (2.5B)	SDXL (6.5B including refiner)	Playground v2.5 (3.5B)	AuraFlow v0.2 (8.7B)	Stable Diffusion 3.5 Large / Large Turbo (8.1B)	FLUX.1 [dev] (12B)	FLUX.1 [schnell] (12B)
NVIDIA GeForce RTX 4060	8	⚠	⚠	⚠	⚠	⚠	⚠	⚠
NVIDIA GeForce RTX 3080	10	✓	⚠	⚠	⚠	⚠	⚠	⚠
NVIDIA GeForce RTX 3060 NVIDIA GeForce RTX 4070 AMD Radeon RX 7700 XT	12	✓	⚠	⚠	⚠	⚠	⚠	⚠
NVIDIA GeForce RTX 4060 Ti NVIDIA GeForce RTX 4070 Ti NVIDIA GeForce RTX 4080 AMD Radeon RX 7800 XT AMD Radeon RX 7600 XT	16	✓	✓	✓	⚠	⚠	⚠	⚠
AMD Radeon RX 7900 XT	20	✓	✓	✓	✓	⚠	⚠	⚠
NVIDIA GeForce RTX 3090 NVIDIA GeForce RTX 4090 AMD Radeon 7900XTX	24	✓	✓	✓	✓	✓	⚠	⚠
NVIDIA H100 AMD Instinct MI250X AMD Instinct MI300A AMD Instinct MI300X	32 (or greater)	✓	✓	✓	✓	✓	✓	✓

Deployment-Tips

- Use messaging system (Service bus) or webhook + progress tracking
- longer timeout settings if you are using restAPI (I had a problem with this myself already)
- consider storing common noise patterns → reload frequently used random noise seeds, reuse initial noise states that produce good results, storing pre computed noise schedule
- different batching strategies, queue management, different queues for different sizes A100/H100 for production
- For low compute power (GPU) use **Quantization**