上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Week 10 Extension: Flow Matching and the DiT Era

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

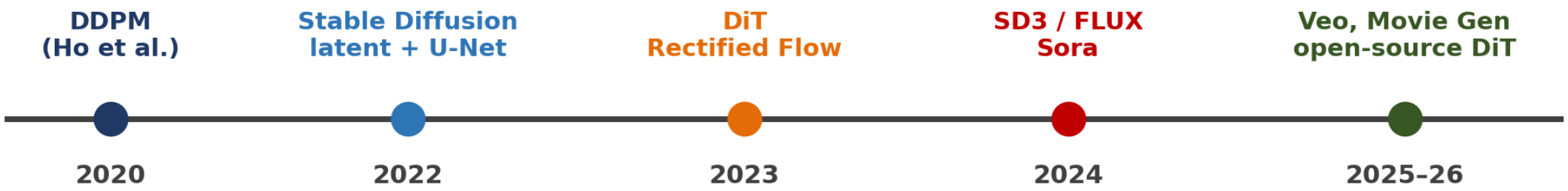
Where L28 Left Off

- L28 closed with the "diffusion broke the trilemma" story — DDPM as the worked example, with flow matching as a 3-slide coda.

Reality check: by 2026, no frontier image or video model is trained as a classical DDPM.

SD3, FLUX, Sora, Movie Gen, Veo — all use flow matching + a Diffusion Transformer (DiT) backbone.

Today: why the field moved, and what the modern recipe actually looks like.



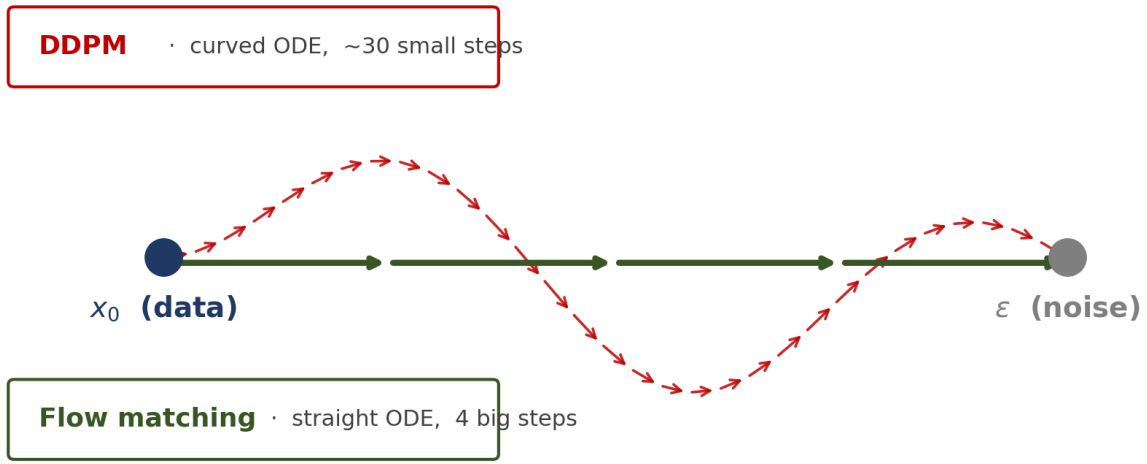
the architecture changed twice (U-Net → DiT, text+image fused → MMDiT), the math changed once (DDPM → flow matching)

From Curved Paths to Straight Lines

- DDPM's reverse process is a curved SDE through noise space — ~50–1000 sampling steps because the denoiser bends along the curve.
- Flow matching reframes it: pick any path between data x_0 and noise ε. The simplest choice — a straight line:

$$x_t = (1-t)\,x_0 + t\,\varepsilon, \qquad t \in [0,1]$$

Straight paths can be integrated with far fewer ODE steps. This is the entire pitch.



Same endpoints — straight paths can be integrated in far fewer ODE steps.

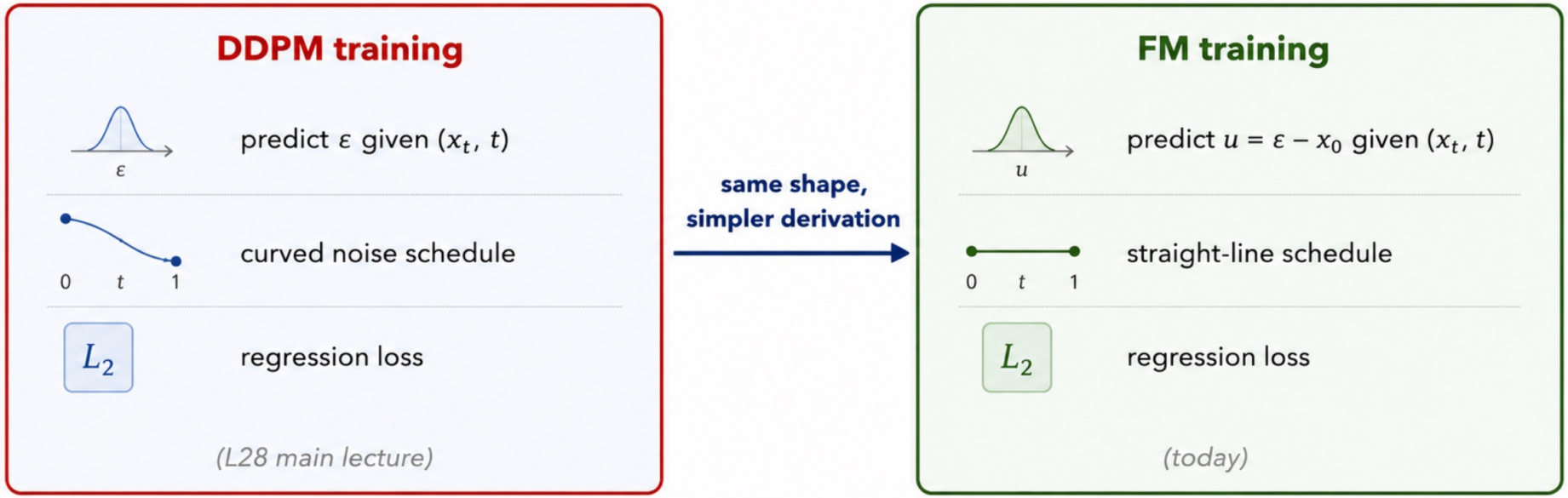
The Flow Matching Recipe

Lipman et al. Flow Matching for Generative Modeling. ICLR 2023

- Define a velocity field v_θ(x_t, t) — at noisy state x_t, which direction toward clean data?
- For the straight-line interpolant, the true velocity is just:

$$u(x_t, t) = \varepsilon - x_0$$

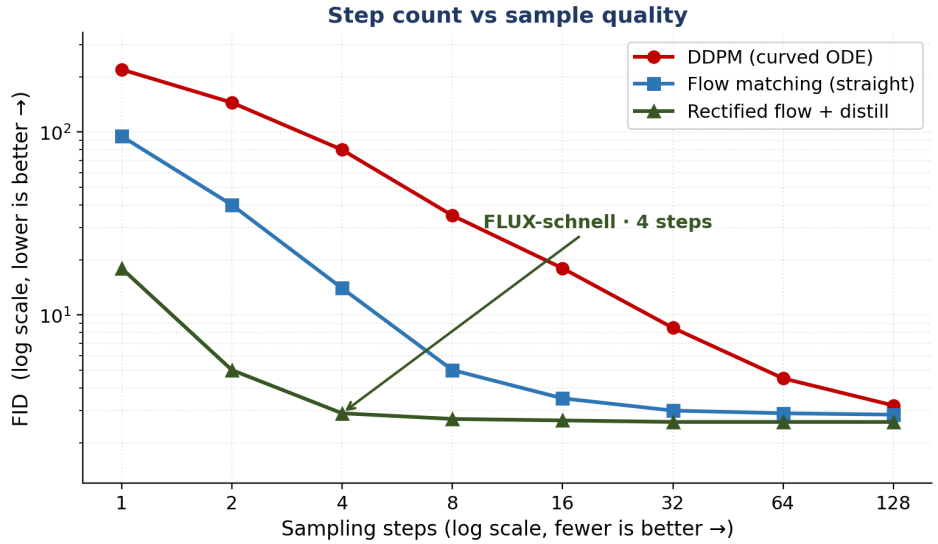
- Train v_θ to regress on this. One MSE loss. No ELBO, no variational machinery.



Rectified Flow and the OT View

Liu et al. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. ICLR 2023

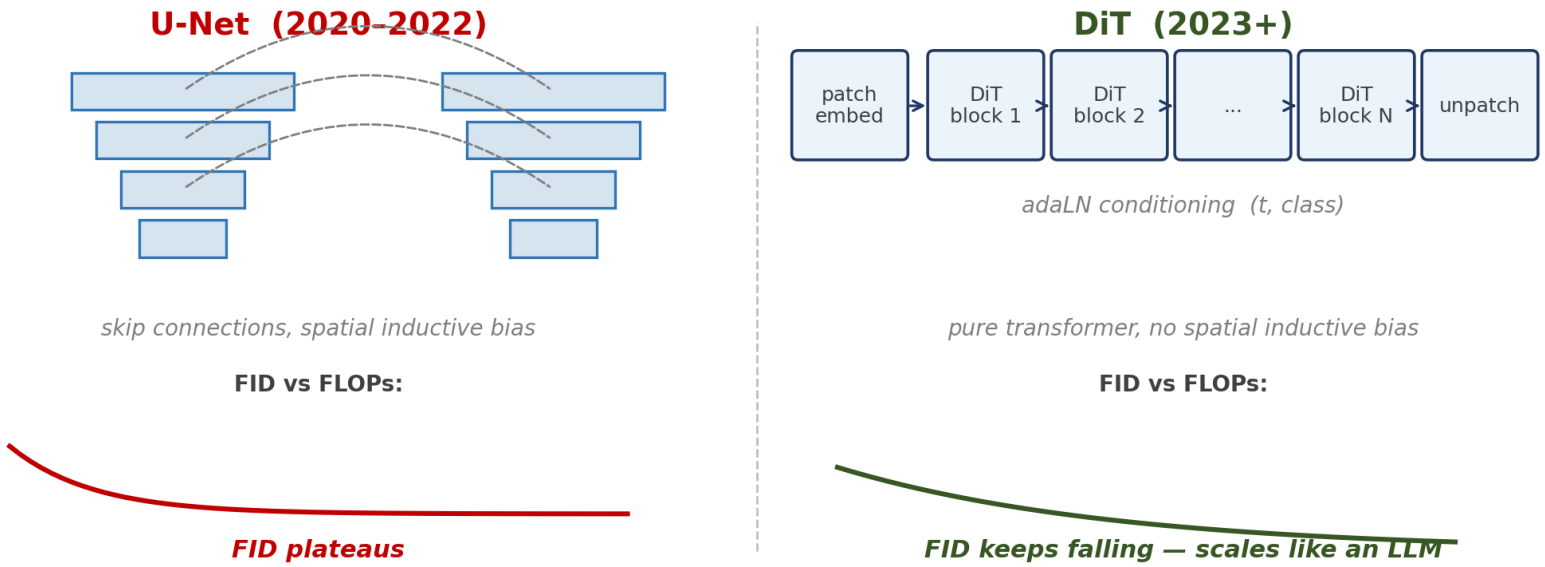
- Same straight-line idea, different motivation: **optimal transport.** Connect data and noise distributions along the straightest coupling.
- **Why straightness matters:** an exactly-straight ODE can be integrated in one Euler step.
- **"Reflow" trick:** train, generate (data, noise) pairs from the model, retrain on those pairs → straighter trajectories each round.
- Powers **few-step models** like FLUX-schnell (4 steps) and SD3.5 Turbo.



DiT: Replacing the U-Net

Peebles & Xie. Scalable Diffusion Models with Transformers (DiT). ICCV 2023

- For 5+ years, every diffusion model used a **U-Net** denoiser — borrowed from segmentation, with strong spatial inductive biases.
- **DiT swaps it out:** tokenize the latent into patches (like ViT), apply a stack of transformer blocks, condition via adaLN.
- **Why it matters:** U-Nets plateaued; DiT scales smoothly with parameters and compute, exactly like an LLM.



MMDiT: The Modern T2I Backbone

Esser et al. Scaling Rectified Flow Transformers (SD3). ICML 2024

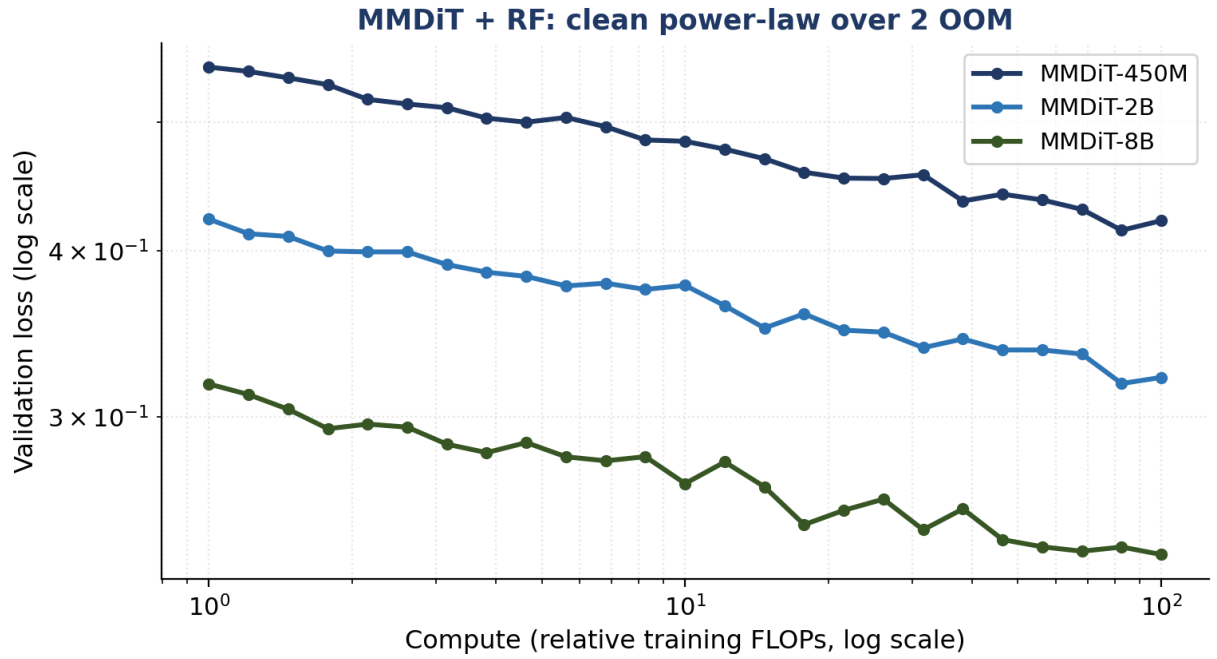
- **Problem:** text and image latents live in very different spaces. A single shared transformer mixes them too aggressively.
- **MMDiT solution:** two parallel transformer streams — text and image — with separate weights but joint attention layers.
- This is the architecture behind **Stable Diffusion 3** and **FLUX.1.**

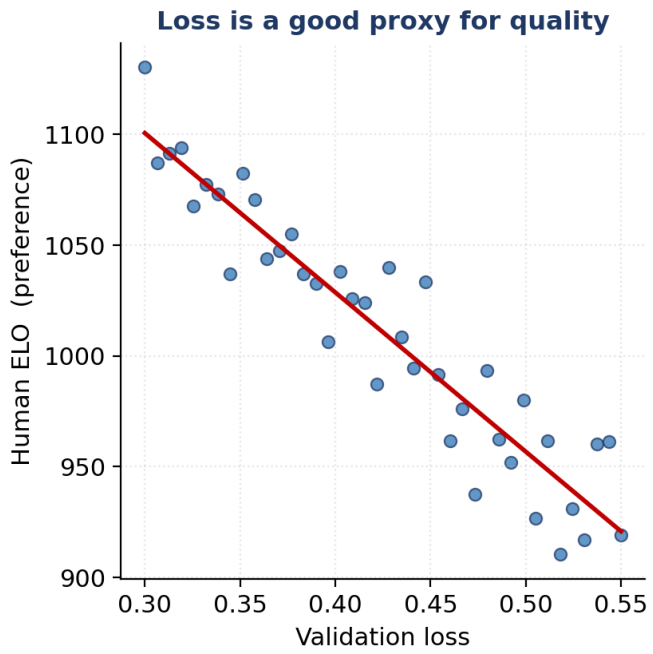
Separate weights per stream · joint attention only where it matters

Scaling Laws for Image Generation

Esser et al. (2024) — second contribution of the SD3 paper

- SD3 trained MMDiT + RF from **450M to 8B parameters** under matched data and compute.
- **Result:** clean power-law scaling — validation loss decreases smoothly, no saturation.
- **Better still:** val loss correlates strongly with GenEval (auto) and human ELO (preference).



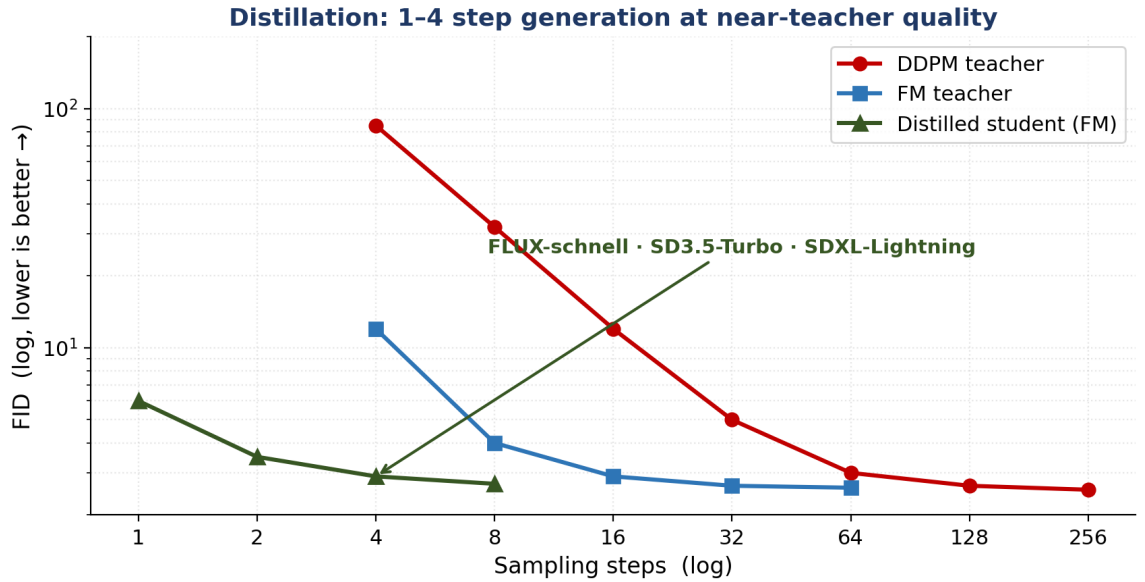


Few-Step Generation: Distillation

Sauer et al. Adversarial Diffusion Distillation. CVPR 2024 / Yin et al. DMD2. NeurIPS 2024

- Even 4 steps is too slow for some products. Goal: **1–4 step generation** at near-teacher quality.
- **Recipe:** distill a slow many-step teacher (FLUX, SD3) into a fast few-step student.
- Adversarial loss + distribution matching (DMD) keep quality high.
- Powers **FLUX.1-schnell, SD3.5 Turbo, SDXL Lightning.** ~10× lower inference cost.

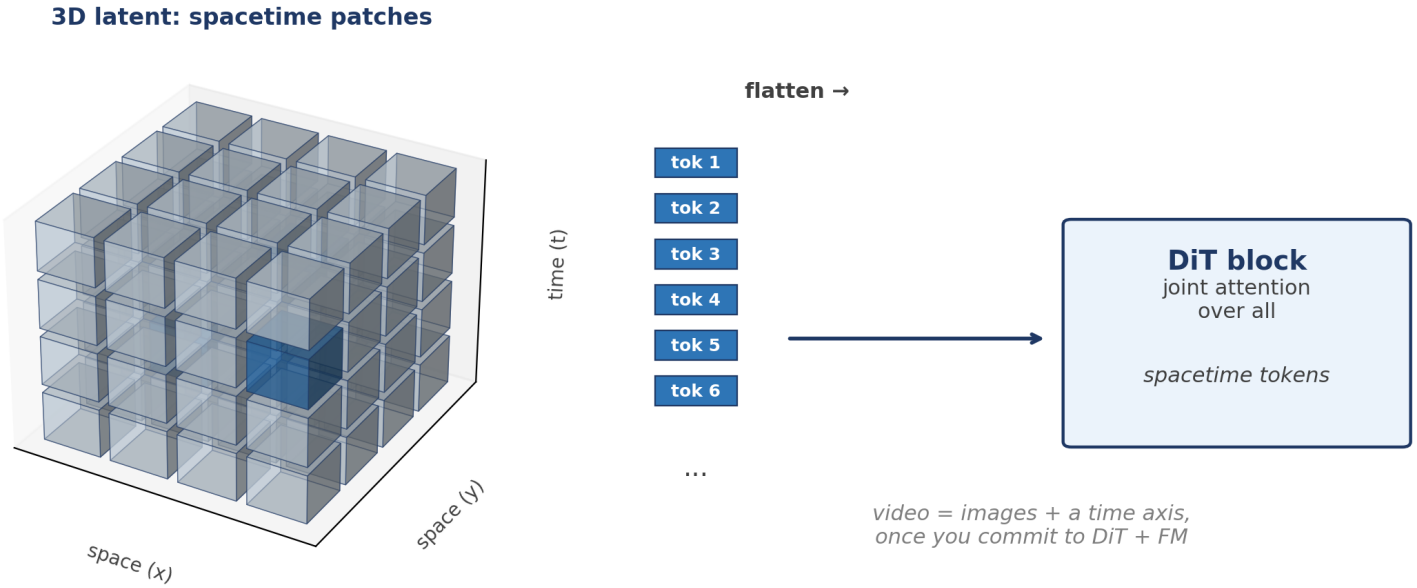
Why it works: straight FM trajectories are easy to compress. Curved DDPM paths are not.



Video: Spacetime Patches

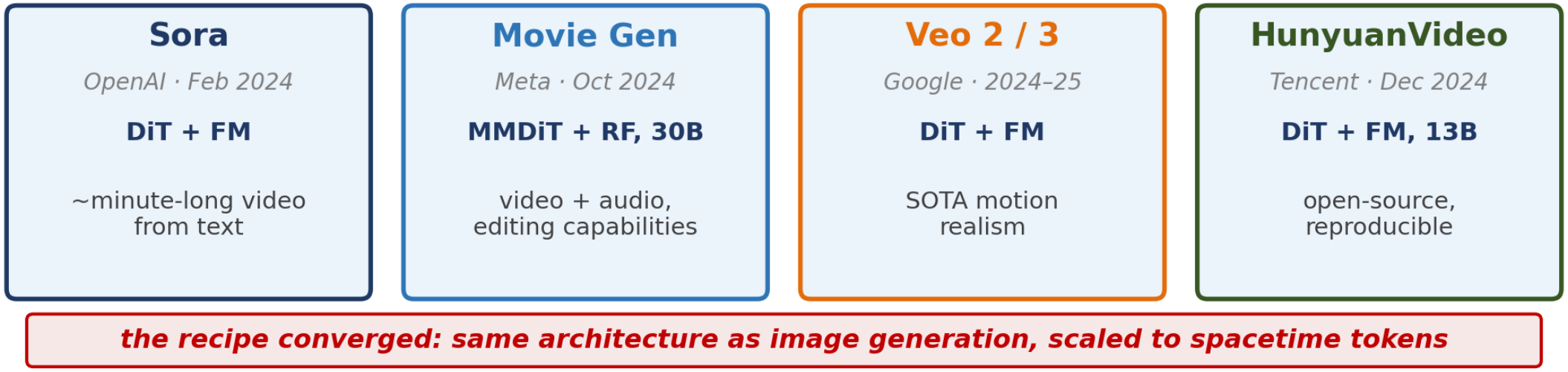
Brooks et al. Sora technical report. OpenAI 2024 / Polyak et al. Movie Gen. Meta 2024

- Video = images + a time axis. The hard part is the tokenizer.
- **Sora's recipe:** encode video → 3D latent (spatial + temporal) → flatten into a sequence of **spacetime patches.**
- From the model's perspective, video is just a longer sequence. Movie Gen / Veo follow the same blueprint.



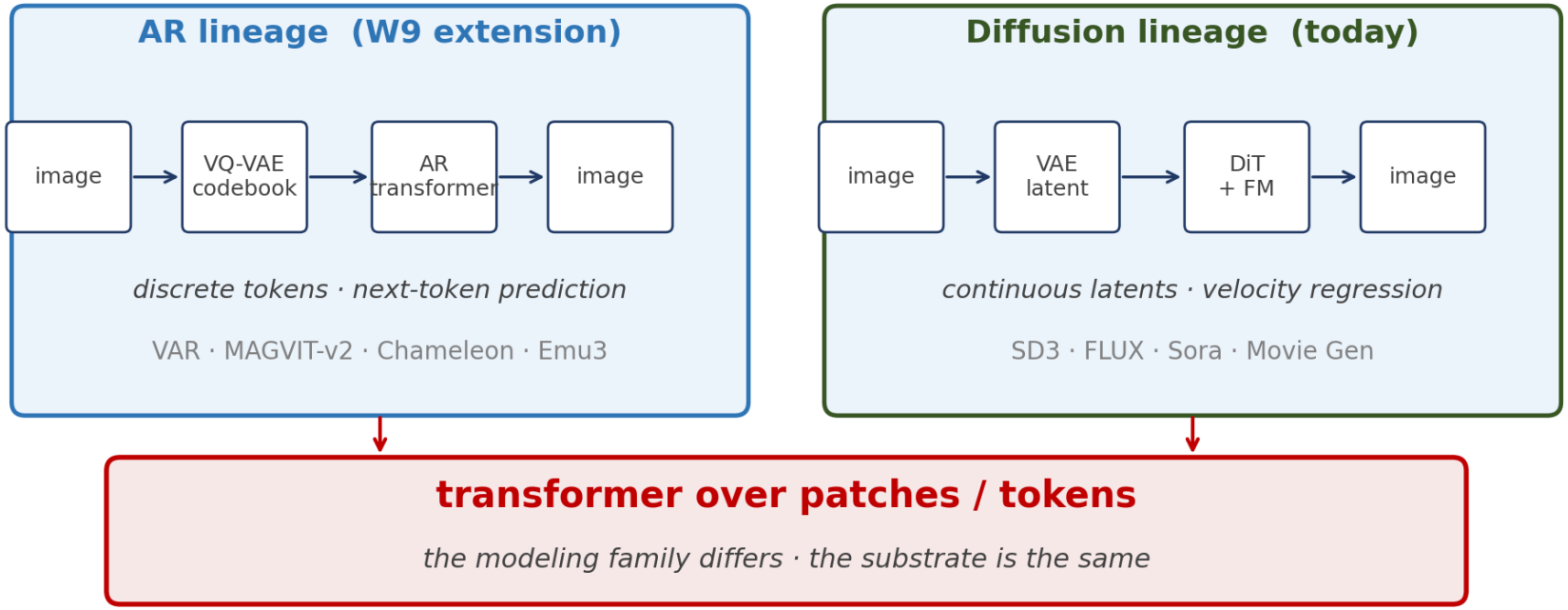
Video: The 2024–26 Lineup

- **Sora** (OpenAI, Feb 2024) — first to demonstrate minute-long, coherent video from a single prompt.
- **Movie Gen** (Meta, Oct 2024) — 30B-parameter video DiT + FM, with audio generation and editing.
- **Veo 2 / Veo 3** (Google, 2024–25) — SOTA on motion realism and instruction following.
- **Open-source:** Mochi-1, HunyuanVideo, Wan-2.1 — DiT + FM is now reproducible outside frontier labs.



Two Lineages, One Architecture

- **AR and diffusion converged on the same substrate.**
- **AR lineage (W9 ext):** discrete tokens (VQ-VAE → LFQ) + transformer next-token prediction.
- **Diffusion lineage (today):** continuous latents + transformer + flow matching.
- **Both run a transformer over a sequence of patches/tokens.** Modeling family and token type differ; the substrate is the same.



Where the Field Went Next

- **AR vs diffusion for video** is still genuinely contested in 2026 — VideoPoet (AR) vs Sora (diffusion) trade wins on different benchmarks.
- **Unified models** do both: *Transfusion* (Meta 2024), *Show-o* (Show Lab 2024), *MAR* (He et al. 2024).
  One transformer trained with FM loss on continuous tokens AND CE loss on discrete tokens.
- **Inference-time scaling** for diffusion — analog of o1-style "think longer" for image gen.
  Best-of-N with a verifier, search over noise initializations. Early results in 2025–26.

Thesis: the modeling family matters less than the architecture and the data. Transformer + scale + clean training objective wins. We've seen this story three times now.

References

Flow Matching & Rectified Flow

[1] Lipman et al. *Flow Matching for Generative Modeling.* ICLR 2023

[2] Liu et al. *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow.* ICLR 2023

Diffusion Transformers

[3] Peebles & Xie. *Scalable Diffusion Models with Transformers (DiT).* ICCV 2023

[4] Esser et al. *Scaling Rectified Flow Transformers for High-Resolution Image Synthesis.* ICML 2024

Distillation

[5] Sauer et al. *Adversarial Diffusion Distillation.* CVPR 2024

[6] Yin et al. *Improved Distribution Matching Distillation (DMD2).* NeurIPS 2024

Video

[7] Brooks et al. *Video generation models as world simulators (Sora technical report).* OpenAI 2024

[8] Polyak et al. *Movie Gen: A Cast of Media Foundation Models.* Meta 2024

Unification

[9] Zhou et al. *Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model.* ICLR 2025

[10] Xie et al. *Show-o: One Single Transformer to Unify Multimodal Understanding and Generation.* ICLR 2025