Flow Matching: From Foundations to Frontiers

A Technical Survey of Five Key Papers (2023–2025)

Theory → OT Coupling → Discrete Extension → Guidance → GM Velocity

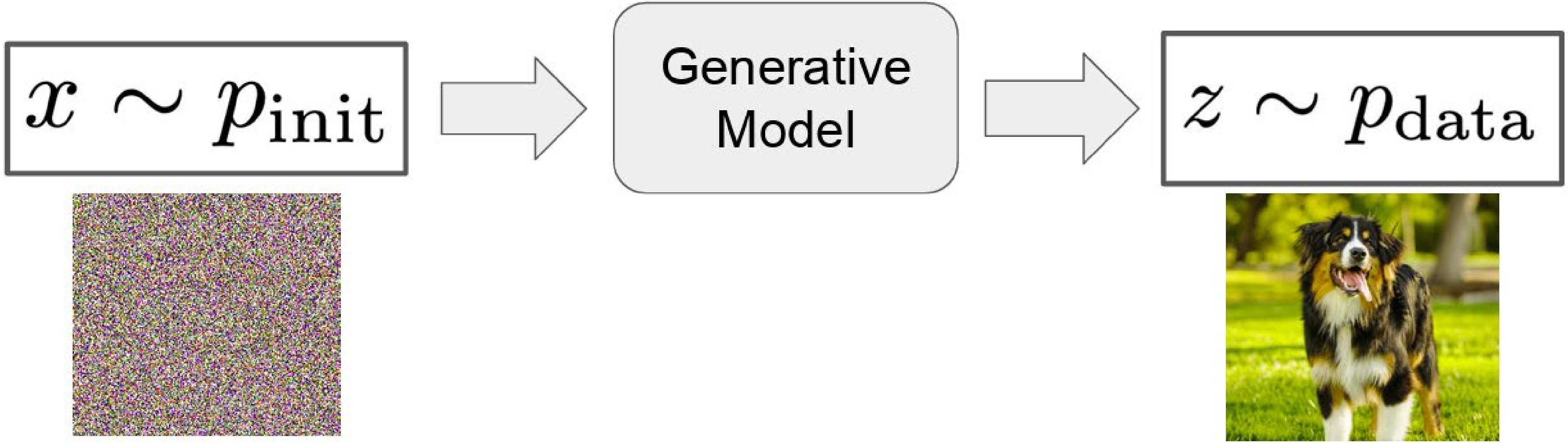
ICLR 2023 • TMLR 2024 • NeurIPS 2024 • ICML 2025 × 2

Discriminative vs. Generative

- Discriminative Models: Learning the Boundary
  - Goal: Model the conditional probability P(Y|X)
  - Function: Maps high-dimensional data X to a low-dimensional regressive/label Y
- Generative Models: Learning the Distribution
  - Goal: model the joint probability P(X, Y) or the data distribution P(X)
  - Function: Maps random noise z~N(0, Id) to valid data X in the target distribution

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

A generative model converts samples from a initial distribution (e.g Gaussian) into samples from the data distribution:



From: https://diffusion.csail.mit.edu/2025/index.html

Diffusion Model

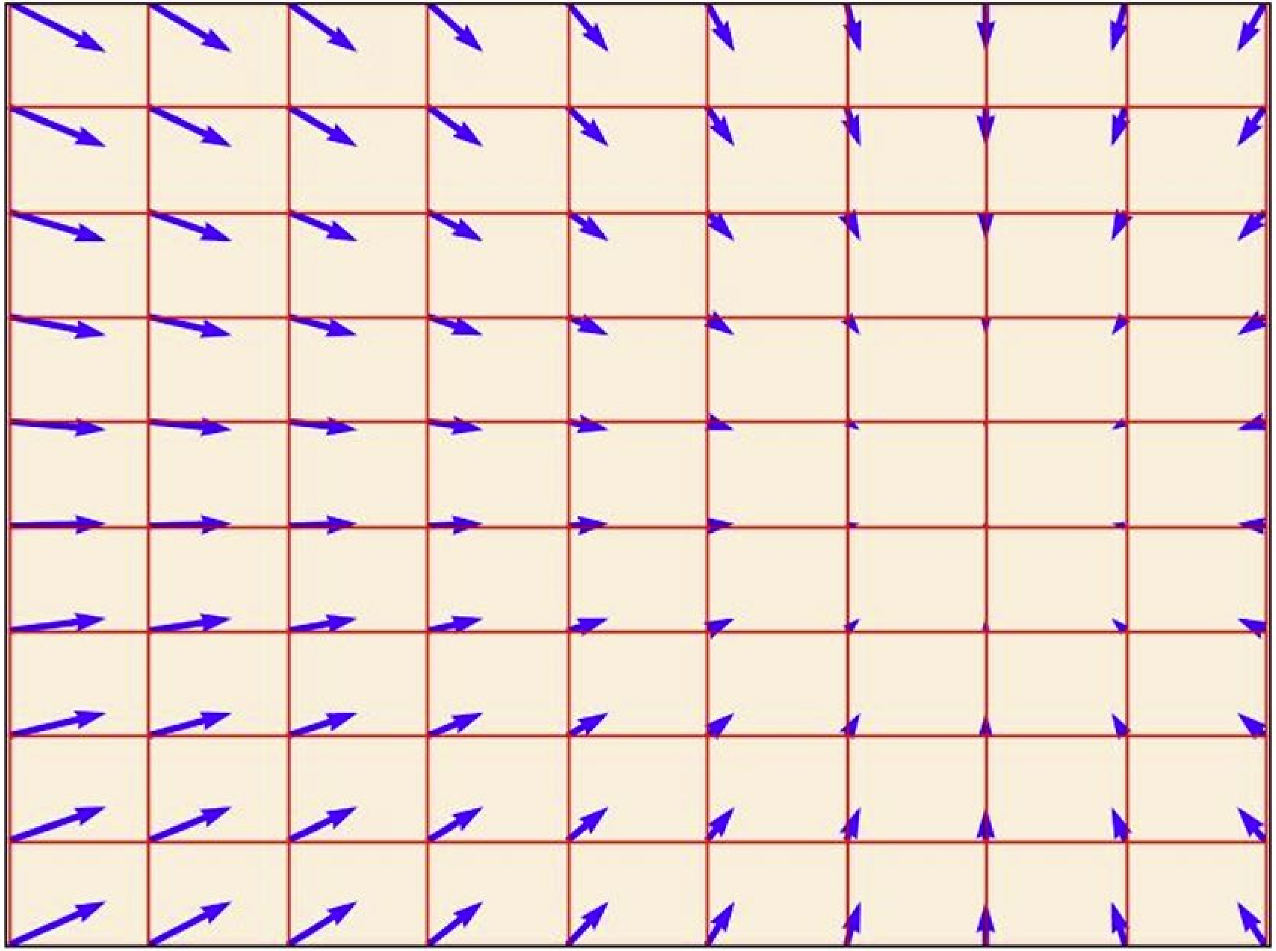
- **The Forward Process (Known Physics):**
  - Gradually destroy data x_0 by injecting Gaussian noise ϵ over T steps.
  - *Formula:* $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$
  - *Result:* x_T becomes pure noise $\mathcal{N}(0, I)$.
- **The Reverse Process (Learned Denoising):**
  - *Goal:* Predict the exact noise ϵ added at step t to subtract it.
  - A neural network (e.g., U-Net) $\epsilon_\theta(x_t, t)$ acts as a step-by-step denoiser.
- **The Training Objective (Surprisingly Simple):**
  - Mean Squared Error between the *actual* noise added and the *predicted* noise:
  - $$\mathcal{L} = \mathbb{E}_{t, x_0, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right]$$
- **The Drawback:**
  - Simulating this reverse stochastic differential equation (SDE) requires hundreds of tiny steps to avoid numerical errors.

- x_0: Original clean data.
- $\epsilon \sim \mathcal{N}(0, I)$: Sampled Gaussian noise.
- $\bar{\alpha}_t \in (0, 1)$: A deterministic decaying schedule.

Enter Flow Matching

- The Core Intuition: Why simulate a "drunkard's walk" (curved, stochastic) when you can draw a straight line?
- The Flow Matching Approach:
  - Shift from predicting noise to learning a continuous Velocity Field $v(x, t)$.
  - Governing ODE: $dx/dt = v_\theta(x, t)$.
- The Straight Path (Optimal Transport intuition):
  - Trajectory: $x_t = (1-t)x_0 + tx_1$.
  - Velocity: $u_t = x_1 - x_0$.
- Advantages:
  - Straight trajectories -> significantly fewer integration steps for generation.
  - General framework: Diffusion models are mathematically just special cases.

Flow - Example



From: https://diffusion.csail.mit.edu/2025/index.html

Roadmap: Five Papers, One Story

1. **Flow Matching for Generative Modeling** — Lipman et al. — ICLR 2023 — Foundation
2. **Minibatch Optimal Transport CFM** — Tong et al. — TMLR 2024 — OT Coupling
3. **Discrete Flow Matching** — Gat et al. — NeurIPS 2024 — Discrete
4. **On the Guidance of Flow Matching** — Feng et al. — ICML 2025 Spotlight — Guidance
5. **Gaussian Mixture Flow Matching** — Chen et al. — ICML 2025 — GM Velocity

Paper 1: Flow Matching for Generative Modeling

Lipman, Chen, Ben-Hamu, Nickel, Le (Meta FAIR) — ICLR 2023

- Background: Continuous Normalizing Flows(CNFs)
  - CNFs are deterministic models defined by an ODE: $\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$.
  - They can model arbitrary probability paths and encompass diffusion processes.
- Motivation:
  - Diffusion models are highly successful but restricted to simple diffusion processes, leading to confined probability paths and slow sampling.
  - We want to train general CNFs with arbitrary probability paths.
- Problem Definition: The Intractability of CNFs
  - Maximum likelihood training of CNFs requires expensive numerical ODE simulations during training.
  - Existing simulation-free methods suffer from intractable integrals or biased gradients.
  - *Goal:* A scalable, simulation-free training algorithm for CNFs.

Proposed Solution (How and Why)

Lipman, Chen, Ben-Hamu, Nickel, Le (Meta FAIR) — ICLR 2023

- The Intractable Goal: Flow Matching(FM)
  - Match a marginal probability path $p_t(x)$ regressing its marginal vector field $u_t(x)$.
  - $\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,p_t(x)} ||v_t(x) - u_t(x)||^2.$
  - Issue: The marginal vector field is intractable to compute directly.
- The Solution: Conditional Flow Matching(CFM)
  - Break the marginal path into simple conditional paths $p_t(x|x_1)$ based on a single target sample x_1
  - Construct a conditional vector field $u_t(x|x_1)$ that generates $p_t(x|x_1)$.
  - New Objective: $\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)} ||v_t(x) - u_t(x|x_1)||^2.$
- Why does this work(Thm 2):
  - Optimizing the tractable CFM objective provides the exact same gradients as the intractable FM objective $\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{CFM}(\theta).$

Technical Depth – Gaussian Paths & Optimal Transport

Lipman, Chen, Ben-Hamu, Nickel, Le (Meta FAIR) — ICLR 2023

- The Universal Gaussian Path:
  - Any path from Noise (x0) to Data (x1) can be defined by a Mean and Variance schedule over time t.
- Unifying Diffusion Models:
  - Standard diffusion uses complex, non-linear schedules (e.g., $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t \epsilon}$).
  - *Insight:* Flow Matching mathematically perfectly replicates these curved paths. Diffusion is just a special case!
- The Breakthrough: Optimal Transport(OT) Path
  - Path: $x_t = (1-t)x_0 + tx_1$.
  - Constant velocity: $u_t = x_1 - x_0$
  - Result: Particles move in absolute straight lines. ODE solvers can take massive step sizes without errors.

Key Contribution

Lipman, Chen, Ben-Hamu, Nickel, Le (Meta FAIR) — ICLR 2023

- A New Training Paradigm (simulation-free CNFs)
  - Proved that learning simple "conditional" paths mathematically equals learning the complex "global" fluid dynamics.
  - *Impact:* We can finally train continuous normalizing flows on massive datasets without expensive ODE solvers.
- The “Grand Unification”:
  - Revealed that all existing Diffusion Models are just specific, sub-optimal instances of Flow Matching with curved paths.
- Optimal Transport Paths:
  - Replaced curved diffusion paths with straight-line, constant-speed trajectories.
  - *Results on ImageNet:* Trains faster, achieves higher image quality (better FID), and requires significantly fewer generation steps than top diffusion models.

Paper 2: Improving CFM with Minibatch Optimal Transport

Tong, Fatras, Malkin, Huguet et al. (Mila / Yoshua Bengio) — TMLR 2024

- Motivation: Paper 1 gave us perfect straight lines for *individual* particles. But what happens to the *global* vector field?
- The Problem: “Independent Coupling”
    - If we randomly pair a noise sample (start) with a data sample (destination), their straight paths will cross each other in space.
    - *The Physics Issue:* At the intersection point, particles are supposed to go in different directions. The velocity field becomes a tangled mess.
    - *The ML Issue:* This causes extreme **variance** in the regression target. The neural network gets confused, making training slow and unstable.

Proposed Solution(How and why)

Tong, Fatras, Malkin, Huguet et al. (Mila / Yoshua Bengio) — TMLR 2024

- **The Goal:** Untangle the trajectories. We want parallel, non-crossing streamlines.
- The Solution: OT Coupling
    - Instead of random pairing, we pair noise and data to minimize the total travel distance of all particles.
- Why it works?
    - By minimizing the total cost, Optimal Transport mathematically guarantees that trajectories will not cross.

Technical Depth(Minibatch Trick) and key contribution

Tong, Fatras, Malkin, Huguet et al. (Mila / Yoshua Bengio) — TMLR 2024

- Minibatch OT
  - The global optimal transport requires O(N^3)
  - Solve OT problem inside small training patch
- The impact
  - **Zero Variance:** As noise approaches zero, the training variance mathematically drops to zero.
  - **Faster Generation:** Because the learned vector field is so smooth and parallel, ODE solvers can take massive steps.
  - **Results:** OT-CFM consistently achieves better image quality (FID) while drastically reducing the Number of Function Evaluations (NFE).

From Continuous to Discrete

Can the flow matching recipe work for text, DNA, and molecules?

Paper 3: Discrete Flow Matching

Gat, Remez, Shaul, Kreuk et al. (Meta FAIR) — NeurIPS 2024 Spotlight

- **Background: Flow Matching works beautifully for continuous data (images, audio)**
  - The velocity field v(x,t) pushes continuous points along smooth trajectories
  - But many critical domains are discrete: text tokens, DNA bases, molecular graphs

- **The Core Challenge:**
  - You cannot "add a velocity vector" to a categorical token — there's no continuous space to move in
  - Token "A" → Token "C" is a jump, not a smooth displacement
  - Prior approaches: embed tokens in continuous space (loses structure) or ad-hoc CTMC formulations

- **The Question This Paper Asks:**
  - Can we build a rigorous discrete analog of the entire Flow Matching framework?
  - Same elegance, same theorems, same training recipe — just for tokens instead of pixels?

Proposed Solution: "Probability Velocity"

Gat et al. — NeurIPS 2024 Spotlight

- **The Key Invention: Replace "spatial velocity" with "probability velocity"**
  - Continuous FM: velocity v(x,t) tells you where to move the point x
  - Discrete FM: probability velocity u_t^i(x^i, z) tells you how fast the probability of token x^i is changing at position i, given current state z

$$u_t^i(x^i, z) \quad \text{for } x^i \in [d], \quad z \in \mathcal{D} = [d]^N$$

- **Constraints (exact analog of CTMC transition rates):**
  - $\sum_{x^i \in [d]} u_t^i(x^i, z) = 0$ — probability is conserved (doesn't appear or vanish)
  - $u_t^i(x^i, z) \geq 0 \quad \text{for } x^i \neq z^i$ — probability flows from current token to others

- **The Discrete Continuity Equation:**

`dp_t(x)/dt + div_x(p_t · u_t) = 0`

$$\frac{d\,p_t(x)}{dt} + \mathrm{div}_x(p_t\, u_t) = 0$$

  - Exact structural analog of the continuous: $\partial_t p_t + \nabla \cdot (p_t\, v_t) = 0$

- **Same Marginalization Theorem holds (Theorem 2):**
  - Train on simple conditional velocities → get the correct marginal velocity for free!
  - Gradient equivalence carries over to discrete setting

Technical Depth: The Structural Parallel

Gat et al. — NeurIPS 2024 Spotlight

- CONDITIONAL PATH (MASK/UNMASK INTERPOLATION)

$$p_t(x^i \mid x_0, x_1) = (1 - \kappa_t)\, \delta_{x_0^i}(x^i) \; + \; \kappa_t\, \delta_{x_1^i}(x^i)$$

  - At time t, each token independently: stays at source (prob 1-κ_t) or jumps to target (prob κ_t)
  - This is "linear interpolation" but in probability space!

- CONTINUOUS FLOW MATCHING (PAPER 1)

$$u_t(x_t) = \frac{\dot{\kappa}_t}{1 - \kappa_t}\left[\hat{x}_{1|t}(x_t) - x_t\right]$$

DISCRETE FLOW MATCHING — x_1-PREDICTION (DENOISER FORM)

$$u_t^i(x^i, z) = \frac{\dot{\kappa}_t}{1 - \kappa_t}\left[p_{1|t}(x^i \mid z) - \delta_{z^i}(x^i)\right]$$

  - Same prefactor κ̇/(1-κ).
  - Point estimate x̂ → probability distribution p.
  - Displacement (x̂ - x) → probability shift (p - δ).

- TRAINING LOSS (CROSS-ENTROPY)

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \mathbb{E}_{t,\ (x_0,x_1)\sim\pi,\ x_t\sim p_{t|0,1}}\left[\log p_{1|t}(x_1^i \mid x_t;\, \theta)\right]$$

  - This is just a masked language model with continuous time t instead of fixed mask ratio

Key Contributions

Gat et al. — NeurIPS 2024 Spotlight

- **Contribution 1: A Rigorous Discrete FM Framework**
  - Complete mathematical parallel: probability velocity, continuity equation, marginalization theorem, gradient equivalence — all carry over
  - Unifies prior discrete methods (D3PM, MDLM, SEDD) as special cases
- **Contribution 2: Corrector Mechanism**
  - Combine forward velocity (unmask) and backward velocity (re-mask) to allow already-decided tokens to be corrected
  - Like Langevin corrector steps in continuous diffusion — but for tokens
- **Contribution 3: Scheduler Design — Cubic $\kappa(t) = t^3$ beats linear $\kappa(t) = t$**
  - Slow early evolution (see global structure first) → fast late refinement (fill in details)
- **Results:**
  - **Language (1.7B params): perplexity 9.7 — approaching Llama-2 7B (8.3), far ahead of same-size AR (22.3)**
  - **Code (HumanEval): 6.7% Pass@1 — first non-autoregressive model with non-trivial coding ability**

Conditional Generation & Guidance

Diffusion guidance relies on score decomposition.

How does guidance work for general flow matching?

Paper 4: On the Guidance of Flow Matching

Feng, Yu, Deng, Hu, Wu (Westlake University) — ICML 2025 Spotlight

- **Background: What is "Guidance" in generative models?**
  - Unconditional generation: model learns p(x) — generates random samples
  - Conditional generation: we want p(x|y) — e.g. generate images matching a text description y
  - Guidance = steering the generation process toward desired properties

- **How Guidance Works in Diffusion Models (well-understood):**
  - Score function decomposes additively:

$$\nabla \log p(x|y) = \nabla \log p(x) + \nabla \log p(y|x)$$

  - Classifier-Free Guidance (CFG): interpolate conditional & unconditional predictions
  - This works because score = gradient of log-density has nice additive properties

- **The Problem with Flow Matching:**
  - FM uses velocity fields, NOT score functions
  - Velocity does NOT decompose additively: $v(x,t|y) \neq v(x,t)$ + something
  - Prior FM guidance only worked for restricted settings (Gaussian source, affine paths)
  - For general FM (OT coupling, non-Gaussian sources from Paper 2) — no guidance theory existed!

Proposed Solution: A General Guidance Framework

Feng et al. — ICML 2025 Spotlight

- **The Goal: Sample from an energy-tilted distribution**
  - Given pretrained FM generating $p(x_1)$, we want:

  $$p'(x_1) \propto p(x_1) \cdot \exp(-J(x_1))$$

  - where $J(x_1)$ is an energy function encoding the desired condition

- **Theorem 3.1 — The Core Result (Exact Guided Velocity):**

  $$v_t'(x) = E_{z\sim p(z|x)} \left[e^{-J(x_1)} / Z_t(x) \cdot u_t(x|z) \right]$$

  - Intuition: at each step, the model looks at all possible couplings $z = (x_0, x_1)$,
  - reweights them by the energy function — paths to "good destinations" get higher weight
  - The guided velocity is this importance-weighted average

- **Why This is Powerful:**
  - No Gaussian source assumption needed
  - No affine path assumption needed
  - No score function decomposition needed
  - Works for ANY Flow Matching model — pure velocity-field-based guidance from first principles

Technical Depth: From General to Classical

Feng et al. — ICML 2025 Spotlight

- **The paper derives a hierarchy of methods under progressive assumptions:**
- **Level 1 — Most General (paper's main contribution):**
  - g^MC (Monte Carlo guidance): sample z~p(z|x_t), compute importance-weighted velocity
  - Training-free, asymptotically exact, works for ANY FM model
  - Cost: requires multiple posterior samples per step
- **Level 2 — Affine paths assumed ($x_t = \alpha_t x_1 + \beta_t x_0$):**
  - Covariance-based guidance: uses Jacobian trick to avoid explicit posterior sampling
  - Faster, but restricted to affine conditional paths
- **Level 3 — Gaussian source + independent coupling (= Diffusion setting):**
  - Gradient approximation: $\nabla_{x_t} J(\hat{x}_1(x_t))$
  - Exactly recovers Classifier-Free Guidance (CFG), DPS, and LGD!
- **The Unification Insight:**
  - **General FM → Affine paths → Gaussian uncoupled → Classical Diffusion Guidance**

Key Contributions

Feng et al. — ICML 2025 Spotlight

- **Contribution 1: First General Guidance Theory for Flow Matching**
  - Theorem 3.1 provides exact guided velocity for arbitrary FM models
  - No restriction on source distribution, path type, or coupling strategy

- **Contribution 2: A Complete Taxonomy of Practical Methods**
  - g^MC (Monte Carlo): exact, training-free, general purpose
  - g^train (Guidance Matching): train a network to predict guidance directly — amortized cost
  - CEP (Contrastive Energy Prediction): contrastive learning approach
  - Gradient approximation: fast, recovers all classical diffusion methods

- **Contribution 3: The "Grand Unification" of Guidance**
  - CFG, DPS, classifier guidance, LGD — all proven to be special cases

- **Experimental Results:**
  - 2D synthetic: g^MC matches exact guided distributions (validates theory)
  - CelebA-HQ 256×256 inverse problems: competitive with specialized diffusion methods
  - MuJoCo offline RL: FM guidance matches diffusion-based planners
  - Demonstrates guidance works beyond images — RL, inverse problems, scientific domains

References

[1] Lipman, Chen, Ben-Hamu, Nickel, Le. "Flow Matching for Generative Modeling." ICLR 2023.

[2] Tong, Fatras, Malkin, Huguet, Zhang, Rector-Brooks, Wolf, Bengio. "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport." TMLR 2024.

[3] Gat, Remez, Shaul, Kreuk, Chen, Synnaeve, Adi, Lipman. "Discrete Flow Matching." NeurIPS 2024 Spotlight.

[4] Feng, Yu, Deng, Hu, Wu. "On the Guidance of Flow Matching." ICML 2025 Spotlight.

[5] Chen, Zhang, Tan, Xu, Luan, Guibas, Wetzstein, Bi. "Gaussian Mixture Flow Matching Models." ICML 2025.

Background:
Song et al. "Score-Based Generative Modeling through SDEs." ICLR 2021.
Ho & Salimans. "Classifier-Free Diffusion Guidance." NeurIPS 2022 Workshop.
Liu et al. "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow." ICLR 2023.
Albergo & Vanden-Eijnden. "Building Normalizing Flows with Stochastic Interpolants." ICLR 2023.