Conditional flow matching for periodic heat and 3D MHD

Race 2026-11-15, team openai · Codex CLI on openai (standard tier, high reasoning) · delivered at 00:51 UTC after 0.33 h · LLM 0.64 USD of 5

Table of contents

1 Question and scope **3**

2 Experiments **4**

3 Evaluation and checks **5**

4 Results **6**

5 Limitations and next work **8**

6 References and AI use **9**

Report of an automated race team (delivered at 00:51 UTC after 0.33 h). Its references to files of the team's repository were rewritten after the race into references to the report's own sections, tables and figures and to the delivered code; the content, numbers and structure are the team's. Delivery: private dataset knoel/milady-20261115-race-openai-delivery on the Hugging Face Hub; the report as the team wrote it is kept with its delivery record, decisions and costs.

1 Question and scope

Can inexpensive physical constraints improve conditional generation of *pairs* of physical fields, and do they preserve distributional quality? We compare an unconstrained flow-matching model, a model trained with soft physics losses, and sampling-time corrections. We first use an analytically solvable periodic heat equation, then a small numerical 3D incompressible MHD system. All numerical results below are from this work, not from the cited papers.

This is a controlled methods study, not a turbulent dynamo simulation. The MHD domain is an unforced periodic box on an 8^3 grid, with random smooth initial conditions and short decaying trajectories. It has no walls, external forcing, sustained dynamo, reversals, realistic geometry, or long-time assimilation experiment. The conditional generator is an experimental input adapter for downstream work, not a validated reduced-order or data-assimilation model.

2 Experiments

Heat. On $[0,2\pi)$, solve $\partial_t q = \kappa \partial_{xx} q$ exactly in Fourier space on 64 points. The initial condition has independent Gaussian sine/cosine coefficients at modes 1–5, with standard deviation $1/k$; the first snapshot is at $t=0.3$ and the second at $t=0.7$. Train on 4,096 independent initial conditions at $\kappa \in \{0.05,0.10,0.20\}$. Test on 256 independent initial conditions per $\kappa = 0.05$, 0.20 (seen) and 0.075, 0.15 (interpolation). The model sees κ and generates both snapshots.

MHD. In Alfvén units with density and permeability one, on $[0,2\pi)^3$, evolve

$$\partial_t u = P(u \times \omega + j \times b) + \nu \Delta u, \quad \partial_t b = \nabla \times (u \times b) + \eta \Delta b, \quad \nabla \cdot u = \nabla \cdot b = 0,$$

where $\omega = \nabla \times u$, $j = \nabla \times b$ and P is the Helmholtz projection. A de-aliased pseudo-spectral solver retains modes $|k_i| \le 2$ and uses RK4 with step 0.05. Initial u and b are independent projected, zero-mean Gaussian random fields, rescaled to random vector RMS ranges $[0.6,1.0]$ and $[0.3,0.7]$. We store pairs at $t=0.50$ and 0.75. The 3,072 training trajectories draw (ν,η) from the 3×3 grid $\{0.04,0.08,0.12\}^2$. The 128 independent test trajectories per condition use (0.04,0.04), (0.12,0.12) and interpolation points (0.06,0.10), (0.10,0.06). Test trajectories are never used for training. Viscosity and resistivity are the control parameters; these values should not be presented as calibrated Reynolds numbers for a real dynamo.

Flow models. For a field pair x and independent standard Gaussian z, regress $v_\theta((1-t)z+tx,t,p)$ against $x-z$ for uniform $t \in [0,1]$, with p the physical parameters. A 3-layer MLP handles heat; a periodic-convolution residual network handles MHD. Both use 2,400 AdamW updates and a fixed 32-step midpoint ODE sampler. We train separate baseline and soft-constraint models with seeds 11, 22 and 33, the same dataset, optimizer budget and sample counts. The heat soft loss adds 0.5 times squared heat-equation residual of the denoised endpoint estimate, in normalized units. MHD adds 0.03 times normalized squared divergence and 0.4 times normalized squared two-snapshot energy-budget residual. Hyperparameters were chosen for this feasibility study; no systematic tuning or early-stopping study was performed.

Sampling corrections. Heat "exact" replaces the second sample by the exact heat evolution of the first at the conditioned κ. MHD "projected" Helmholtz-projects each vector and removes modes outside $|k_i| \le 2$. "Projected + budget" further scales both second-snapshot vectors by the same nonnegative factor. For $E = \tfrac{1}{2}\langle |u|^2+|b|^2 \rangle$ and $D = \nu\langle |\nabla u|^2 \rangle + \eta \langle |\nabla b|^2 \rangle$, the factor enforces the trapezoidal discrete relation $E_1 - E_0 + \Delta t(D_0+D_1)/2 = 0$, with $\Delta t = 0.25$. These are deliberately strong interventions: exact heat and the budget residual approach zero by construction, not because the generative model learned the PDE. The MHD correction does not enforce the full momentum or induction equations, and can change the target distribution.

3 Evaluation and checks

For each condition, a separate 128-trajectory MHD or 256-trajectory heat reference sample is compared with an equally sized generated sample. One-dimensional energy and helicity distribution errors are empirical Wasserstein-1 distances; spectral errors are L1 differences of mean shell energies. We also calculate reference-versus-another-independent-reference distances, which expose finite-sample variation rather than acting as a zero-error target. Model comparisons share sample noise within a trained model. Seen and interpolated conditions are averaged separately; seed ranges describe only three trainings. Parameter intervention holds the simulator's initial random draw or the generator's latent noise fixed while changing the parameter.

MHD diagnostics include total, kinetic, and magnetic energy distributions; kinetic and magnetic energy shell spectra; full-frequency Fourier divergence RMS averaged across the four vectors; absolute trapezoidal energy-budget residual; magnetic helicity of the zero-mean solenoidal fluctuation in Coulomb gauge; current helicity; and a topology proxy. The proxy counts 6-connected components of the |b| excursion set above each field's 75th percentile, with periodic wrapping. It measures a coarse geometric feature, not field-line linkage or topology preservation. Helicity for unconstrained, non-solenoidal fields is only a proxy. The input-grid Nyquist convention in training sets the real derivative's Nyquist coefficient to zero; the reporting evaluator counts all signed Fourier frequencies in divergence, dissipation and spectra. The hard projector removes those high modes. Current helicity uses the real-grid curl convention. The reporting evaluator is part of the delivered code.

Independent tests in the delivered code check projection idempotence/divergence, the spectral MHD energy identity, an exact heat Fourier mode, and the correction's discrete budget. For one double-precision de-aliased trajectory, relative L2 errors against step 0.0125 were 1.86×10^{-6}, 1.12×10^{-7}, and 6.47×10^{-9} for steps 0.1, 0.05, and 0.025 respectively (convergence results in the delivery), consistent with temporal fourth-order convergence in this check. This is not a spatial resolution study.

4 Results

The table summarizes the two interpolation conditions, averaged over conditions and the three training seeds. Brackets are the range of the three seed-level means, not confidence intervals. Full seen-condition results, all metrics and per-seed values are in the results summary and the corrected JSON and per-sample arrays of the delivery dataset on the Hugging Face Hub.

Heat strategy	PDE residual RMS	energy W1	spectrum L1
Baseline	0.487 [0.476, 0.498]	0.159 [0.145, 0.182]	0.188 [0.177, 0.208]
Soft loss	0.266 [0.265, 0.267]	0.098 [0.082, 0.120]	0.113 [0.099, 0.128]
Baseline + exact evolution	0 by construction	0.070 [0.052, 0.091]	0.083 [0.062, 0.109]
Independent reference vs reference	0	0.070	0.047

The heat soft loss improved both residual and the two distributional errors in this experiment. The exact-evolution correction also improved these metrics, but its zero residual is guaranteed by construction. With matched noise, raising κ from 0.05 to 0.20 changed mean second-snapshot energy by −0.1775 in the simulator, versus −0.041 to −0.058 for raw baseline models and −0.149 to −0.174 after exact evolution. Thus the raw generator's parameter response was too weak on this intervention.

MHD strategy	divergence RMS	budget residual	total-energy W1	magnetic spectrum L1	component-count W1
Baseline	0.178 [0.176, 0.182]	0.00904 [0.00825, 0.01036]	0.0358 [0.0349, 0.0374]	0.0151 [0.0151, 0.0152]	1.62 [1.34, 2.00]
Soft loss	0.209 [0.168, 0.289]	0.00998 [0.00836, 0.01312]	0.0426 [0.0343, 0.0590]	0.0185 [0.0144, 0.0260]	2.20 [1.65, 3.19]
Baseline + projection	1.18×10^{-7}	0.00676	0.0344	0.0131	0.672

MHD strategy	divergence RMS	budget residual	total-energy W1	magnetic spectrum L1	component-count W1
Baseline + projection + budget	1.19×10^{-7}	1.80×10^{-8}	0.0337	0.0145	0.672
Independent reference vs reference	1.47×10^{-7}	0.000412	0.00788	0.00362	0.590

In MHD, hard projection reduces divergence to numerical precision and the added energy correction enforces its chosen discrete budget; neither proves the full PDE. The soft loss was inconsistent across seeds and did not beat the baseline on these aggregate metrics. The projected-plus-budget energy and magnetic-spectrum discrepancies remain well above the reference/reference comparison; the budget correction slightly *increased* the magnetic spectral error relative to projection alone. Kinetic spectrum L1 for the projected-plus-budget baseline was 0.0217 versus reference/reference 0.00535. Magnetic-helicity W1 was 0.00183 versus reference/reference 0.00179; this small-sample comparison does not establish topology preservation. The component-count metric is only a coarse proxy and also depends on resolution.

For matched MHD initial conditions, increasing (ν,η) from (0.04,0.04) to (0.12,0.12) changed mean total energy by -0.1475; raw baseline seed means were -0.0349 to -0.0490, and projection-plus-budget means were -0.0846 to -0.1025. The sign is correct but response magnitude remains weak. The fixed-noise coupling is a sensitivity diagnostic, not proof that individual generated pairs represent the simulator's corresponding trajectories.

There was no formal hypothesis test: three training seeds are inadequate for reliable population-level significance, and bootstrap uncertainty over the shared finite test set would not resolve training-seed uncertainty. Reference/reference distances, seed ranges and all raw sample-level metrics are supplied for inspection.

5 Limitations and next work

The small, finite and non-turbulent ensemble, low spatial resolution, single domain/initial-condition family, fixed two-time horizon, only three model seeds, and absence of external real MHD data limit generalization. The interpolation test is within the range of training coefficients; extrapolation, parameter-dependent forcing and boundary conditions were not tested. We did not compute field-line mapping, magnetic-linking invariants, multi-time dynamics, reversals, or downstream assimilation skill. A low residual alone is insufficient evidence of a realistic distribution. Follow-up should use independently validated higher-resolution forced dynamo runs, held-out regimes and geometry, genuine field-line topology, temporal rollout tests, downstream ROM/assimilation evaluation, and independently calibrated uncertainty.

6 References and AI use

- Lipman, Chen, Ben-Hamu, Nickel and Le, “Flow Matching for Generative Modeling,” ICLR 2023, arXiv:2210.02747. Our conditional linear-path training follows the general flow-matching idea; this work is not a reproduction of their published results.
- Bastek, Sun and Kochmann, “Physics-Informed Diffusion Models,” ICLR 2025, arXiv:2403.14404. Physics-based training losses motivate the comparison; their reported results are not our measurements.
- The subject brief is the internship offer.

OpenAI GPT-6.1-terra, operated through Codex, designed and implemented the study, ran and interpreted the experiments, and wrote the report. No other AI model contributed research content. Launcher connectivity probes of other models are recorded in the framework ledger. The numerical PDE and generative results are computed by the committed code, not generated by the language model. No human validation or MilaDy toolbox maintainer review occurred.