Physics-Constrained Conditional Flow Matching for Parametric Dynamical Systems

Race 2026-11-15, team zhipu · ZCode on zhipu (standard tier, high reasoning) · delivered at 01:53 UTC after 1.36 h · LLM 2.33 USD of 5

Table of contents

0.1 A comparative study on Kolmogorov flow and solenoidal helical fields 2

1 1 Introduction and problem statement 3

2 2 Related work 4

3 3 Methods 5
3.1 3.1 Conditional flow matching 5
3.2 3.2 2D Kolmogorov-flow benchmark and the data-level constraint . . 5
3.3 3.3 Sampling-level constraints 5
3.4 3.4 3D solenoidal helical benchmark and the four constraint strategies 6
3.5 3.5 Metrics 6

4 4 Experimental setup 7

5 5 Results 8
5.1 5.1 Benchmark data quality (physics constraints at the data level) . . 8
5.2 5.2 2D: conditional flow matching and sampling-time constraints . . . 8
5.3 5.3 3D: training the four strategies 9

6 6 Discussion and limitations 10

7 References 11

8 Author contributions (AI models) 12

Report of an automated race team (delivered at 01:53 UTC after 1.36 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-zhipu-delivery on the Hugging Face Hub; the report as the team wrote it is kept with its delivery record, decisions and costs.

0.1 A comparative study on Kolmogorov flow and solenoidal helical fields

Mission race-zhipu — ANR MilaDy internship subject (stage de fin d'étude 2027), compact empirical study. Date: 2026-11-15. Authorship: this study was produced end-to-end by AI models (see *Author contributions*).

---

1 1 Introduction and problem statement

Magnetohydrodynamic (MHD) dynamos — the self-excited generation of magnetic fields by the turbulent motion of electrically conducting fluids in planetary and stellar cores — are a canonical hard problem for data-driven modelling. Their magnetic fields are subject to strong, non-negotiable physical constraints: the field is divergence-free ($\nabla \cdot B = 0$) everywhere and at all times; its energy balance is governed by the competition between induction, advection, and Ohmic dissipation, so the statistical steady state is a delicate function of the control parameters; and the system is sharply parameter-sensitive — the magnetic Reynolds number, the fluid Reynolds number Re, and the Prandtl numbers control whether a dynamo sustains, saturates, or decays, and whether it undergoes stochastic polarity reversals on very long timescales. Direct numerical simulation (DNS) of these regimes is expensive, and exploring the parameter space or generating statistically representative ensembles is often the dominant cost.

This motivates physics-constrained generative models: learned conditional samplers of the map from parameters to physical fields that can emulate the parameter dependence of the underlying PDE without resolving its full dynamics, while still respecting the constraints the governing equations and boundary conditions impose. The ANR MilaDy internship subject asks precisely for this — conditional generative models (diffusion or flow matching) that generate physical fields conditioned on system parameters, respecting constraints derived from the governing PDEs and boundary conditions, validated first on simple PDE benchmarks, and then evaluated with statistical *and* physical consistency metrics (energy spectra, helicity, divergence control, topology preservation), with MHD dynamos as the motivating application.

Our scope, given a compact compute budget, is deliberately narrower than full MHD. We treat the problem as a two-stage benchmark: (i) forced two-dimensional Navier–Stokes ("Kolmogorov flow"), where incompressibility is imposed exactly at the data level, and (ii) a synthetic three-dimensional solenoidal helical field benchmark built in Fourier space, which reproduces the statistical structure of divergence-free helical magnetic fields and stands in for 3D MHD magnetic fields (a 3D MHD DNS is beyond the available compute). Throughout, the guiding question is *where* the physics constraint should enter — in the data/field formulation, in the model architecture, in the training loss, or in the sampling procedure — and what each choice buys, and costs, in fidelity and speed.

2 2 Related work

Several recent lines of work frame our contribution. Wang et al. (2026, *Fundiff*) introduce diffusion models over function spaces for physics-informed generative modeling, arguing that continuous function-space formulations decouple the generative model from a fixed discretization. Fukami and Taira (2023) show that extreme aerodynamic flows can be grasped on a low-dimensional manifold, motivating surrogate models that operate on compact learned representations rather than raw grids. Bansal et al. (2024, ICLR) develop *universal guidance* for diffusion models, a unified framework in which arbitrary constraints (including physics-based ones) are imposed at sampling time through a guidance term rather than baked into training. Bastek, Sun, and Kochmann (2025, ICLR) embed physical knowledge directly into diffusion models, enforcing governing-equation structure during training and generation. Solera-Rico et al. (2024, *Nature Communications*) combine a β-VAE with transformers for reduced-order modelling of fluid flows, illustrating the data-driven reduced-modelling route to cheap flow surrogates.

These works differ mainly in *where* physics enters: at the data/formulation level (state representation, e.g. working in a divergence-free basis), at the model-design level (architectures that are divergence-free by construction), at the training-loss level (soft penalties), or at the sampling level (guidance or post-hoc projection). Each is usually demonstrated in isolation. Our contribution is a controlled, comparative study across exactly these four injection points, evaluated on a common conditional generative task and scored with the same physical and statistical metrics. We do not claim to outperform any single method; we aim to make the trade-off between fidelity, physical consistency, and cost legible.

3 3 Methods

3.1 3.1 Conditional flow matching

We use conditional flow matching with the rectified-flow path: for a data sample x1 and a noise draw x0 ~ N(0, I), the interpolation is x_t = (1−t) x0 + t x1, for t ∈ [0,1], whose marginal velocity field is v = x1 − x0. A neural network v_θ(x_t, t, c) is trained by regression on this target velocity, with conditioning c injected via FiLM layers. Sampling is the ODE dx_t/dt = v_θ(x_t, t, c), integrated from x0 ~ N(0, I) with an Euler scheme. All constraints described below modify either the target distribution, the architecture, the loss, or the ODE drift at sampling, leaving the core matching objective otherwise unchanged.

3.2 3.2 2D Kolmogorov-flow benchmark and the data-level constraint

The first benchmark is forced two-dimensional Navier–Stokes on a 64×64 doubly periodic grid, integrated with a pseudo-spectral solver (second-order Runge–Kutta, fp32, 2/3 dealiasing). The Kolmogorov forcing is f = (A sin(2π kf y), 0) with A = 1 and kf = 4, and the control parameter is the Reynolds number Re = A/(ν kf), varied through the viscosity ν ∈ {0.02, 0.01, 0.006, 0.004} for training, with ν = 0.008 (Re = 31.25) held out for interpolation testing. Incompressibility is imposed *at the data level*: the state is the scalar vorticity ω, and the velocity is recovered by the streamfunction/Biot–Savart mapping, so ∇·u = 0 identically for every generated field, by construction, without any architectural or sampling machinery.

3.3 3.3 Sampling-level constraints

Two sampling-time constraints are tested on the 2D benchmark. (a) *proj*: each sampled field is spectrally re-projected onto the Reynolds-conditioned training energy spectrum S_train(k; Re) — a cheap, data-informed post-processing that enforces the correct spectral content. (b) *guided*: normalized-gradient *energy guidance*, which adds to the ODE drift a normalized gradient step (strength η = 0.05 per step) that pulls the energy 0.5⟨|u|²⟩ of the implied terminal field (x̂1 = x_t + (1−t) v_θ) toward the training energy at that Re, with the target interpolated for the held-out Re. Guidance thus biases the trajectory toward energy-consistent endpoints without retraining.

3.4 3.4 3D solenoidal helical benchmark and the four constraint strategies

The second benchmark stands in for 3D MHD magnetic fields without a full MHD DNS. Synthetic solenoidal helical random fields are constructed in Fourier space from Beltrami (helical) modes $e\pm$ satisfying $i\,k \times e\pm = \pm|k|\,e\pm$. Two parameters $\theta = (kp, h)$ define the family: $kp \in [2.5, 5.5]$ controls the shell-energy profile $g(k) = \exp(-((k-kp)/1.5)^2)$, and $h \in [0, 0.9]$ is the fraction of energy placed in positive-helicity modes; each half-space mode's variance is normalized so the total energy $0.5\langle|B|^2\rangle = 1$ exactly. The fields are exactly divergence-free by construction. Four strategies are compared: **S0** raw conditional flow matching (baseline, no constraint); **S1** base + *proj*: a hard, data-free Fourier projection onto the divergence-free subspace, $\hat{B} \to \hat{B} - k(k\cdot\hat{B})/|k|^2$, applied at sampling; **S2** *curl*: a divergence-free-by-design architecture in which flow matching is performed in vector-potential space with Coulomb-gauge targets $A = i\,k \times \hat{B}/|k|^2$ and the sampled field is $B = \nabla\times A$; and **S3** *divloss*: raw flow matching plus a divergence penalty $\lambda\cdot\langle \mathrm{div}(\hat{x}1)^2\rangle/\langle|\nabla\hat{x}1|^2\rangle$ added to the training loss (a soft training-time constraint). S1 and S3 leave the data and architecture untouched but enforce the constraint at sampling or in the loss, respectively; S2 moves the constraint into the model design.

3.5 3.5 Metrics

We score each model on statistical and physical consistency. Statistical: the radial shell spectrum $S(k)$ with relative L2 error against the reference; the energy error; and RBF-MMD^2 between sampled and reference distributions of standardized summary statistics (log-spectrum, rms, kurtosis or helicity ratio, and Beltrami alignment). Physical: the divergence ratio $\mathrm{sqrt}(\langle \mathrm{div}^2\rangle/\langle|\nabla B|^2\rangle)$; the helicity ratio $\langle B\cdot\nabla\times B\rangle/\langle|B|^2\rangle$; the Beltrami alignment $|\langle B\cdot\nabla\times B\rangle|/(\langle|B|^2\rangle\langle|\nabla\times B|^2\rangle)\text{^}0.5$; and the Q-fraction, the fraction of cells with $Q = \omega^2 - |S|^2 > 0$ (a flow-topology statistic discriminating vorticity-dominated from strain-dominated regions). Sampling throughput (fields per second) is reported to expose the computational cost of each constraint strategy. All spectral and differential operators are computed in Fourier space, so the divergence diagnostics are exact up to floating-point round-off (~1e-7 in our normalized units).

4 4 Experimental setup

2D: 64×64 periodic grid; four training viscosities $\nu \in \{0.02, 0.01, 0.006, 0.004\}$ (Re = 12.5, 25, 41.7, 62.5) plus $\nu = 0.008$ (Re = 31.25) held out; 50 snapshots per ν (dt = 1e-4, T = 10, transient 2.5, snapshot every 0.15), 250 in total, with an 80/20 temporal split for the training ν. The solver is pseudo-spectral (fp32, RK2, integer-wavenumber 2/3 dealiasing, advective-CFL-safe dt). 3D: 32^3 periodic grid; 2000 training fields with random θ in the box, plus 4 fixed test parameter sets θ × 200 fields each. Both conditional U-Nets are small (1.0 M / 1.9 M parameters) with FiLM conditioning on time and the control parameter (Re or (kp, h)); the 2D model is trained for 1500 steps (batch 64) and 3D models for 1200 steps (batch 16, the three 3D runs executed concurrently on one GPU), AdamW lr 2e-4. Sampling uses a 50-step Euler ODE solver with 64 generated fields per parameter point. All runs use a single NVIDIA RTX 4090 (24 GB, rented on Runpod) in fp32 with fixed seeds (data seed 0, model seed 0). Every number in §5 is reproducible from the run script of the delivered code on one CUDA GPU (~25 GPU-minutes); the 3D dataset generator is pure NumPy and regenerates byte-identical data for a given seed. Honest framing: the 3D set is a synthetic benchmark, not MHD DNS; the Re range is moderate (no inertial-range claims); the models are small; and we report a single seed per configuration.

5 5 Results

5.1 5.1 Benchmark data quality (physics constraints at the data level)

The 2D dataset (250 snapshots, 50 per ν) is finite and statistically ordered: the mean training energy $0.5\langle|u|^2\rangle$ grows from 1.8×10^{-3} (Re = 12.5, near-laminar) through 3.3×10^{-2} (Re = 25) and 0.46 (Re = 41.7) to 3.11 (Re = 62.5, strongly turbulent) — matching the laminar scaling $A^2/(2\nu^2(2\pi k_f)^4)$ at low Re and an inverse-cascade-dominated regime at high Re. The radial vorticity spectra show the Kolmogorov forcing ring at k = k_f = 4 (peak shell value up to 4×10^6 at Re = 25) superposed on a large-scale k = 1 mode that dominates at Re = 62.5. The 3D dataset passes every constraint check exactly: total energy $0.5\langle|B|^2\rangle = 1.009 \pm 0.044$ (target 1, N^6-Parseval normalization), Fourier-exact divergence ratio $\leq 1.5\times10^{-7}$ for every field, mean helicity ratio $\langle B\cdot\nabla\times B\rangle/\langle|B|^2\rangle = 2.50$ and Beltrami alignment 0.47.

5.2 5.2 2D: conditional flow matching and sampling-time constraints

All three 2D strategies were trained on identical data (fm2d, 1500 steps, final loss 0.051) and evaluated by 50-step ODE sampling of 64 fields per Re point. Headline metrics (means over the 5 Re points; full per-Re tables regenerate via the run script of the delivered code):

strategy	spectrum rel. L2 error ↓	MMD^2 (summary stats)
base (plain CFM)	1.31	0.82
base + guided (energy guidance $\eta = 0.05$)	1.81	0.94
base + proj (spectrum re-projection)	2.27	1.23

A direct probe of the trained model (64 samples per point, spectra against the reference split) explains these numbers. The *data* spectra are strongly regime-dependent — forcing-ring-dominated at low Re ($S(k=4) = 4.0\times10^6$ at Re = 25) and large-scale-dominated at Re = 62.5 ($S(k=1) = 1.4\times10^8$) — but the underfit CFM (1.0 M parameters, 1500 steps, 200 training fields) produces a blurred mixture of the two regimes: it over-weights the k = 1 shell at low Re and under-resolves the k = k_f ring, with per-point energy errors of +185 % at Re = 25 and +172 % at Re = 62.5, and −71 % at the held-out Re = 31.25 (energy 0.028 vs 0.095). Sampling-time constraints cannot repair an underfit *conditioning*: energy guidance corrects the per-point energy bias but leaves the spectral misallocation intact, and the hard spectrum re-projection constrains only the one-point spectrum while the phases (and hence the MMD statistics) remain those of the base model. The

honest conclusion for this benchmark is negative and instructive: **at this training budget, sampling-level constraint enforcement does not substitute for a well-trained conditional model**; the constraint-injection hierarchy must be compared on models whose conditional fidelity is not the bottleneck.

5.3 5.3 3D: training the four strategies

The three 3D models trained to convergence within the budget (1200 steps, final FM losses: base 0.390, curl 0.020, divloss 0.408). The curl model's much lower loss reflects its target space: in A-space the fields are smoother ($A = i\,k \times \hat{B}/k^2$ attenuates each mode by $1/k^2$), not a better model of B. The divloss model's divergence diagnostic on the implied terminal fields $\hat{x}_1$ remained 0.40 (vs 1.5×10^{-7} for data) at $\lambda = 0.1$ — the soft training-time penalty did **not** enforce the constraint at this weight, confirming that soft penalties need careful weighting/schedules that a fixed-budget study cannot tune.

The full 3D sampling/evaluation stage (4 strategies × 4 parameter points × 64 fields with spectrum, divergence, helicity, alignment and MMD metrics) was still executing on the GPU when the mission's delivery deadline was reached; its metrics file is produced by step 4 of the run script of the delivered code and regenerates in ≈ 8 minutes. The delivered artifacts therefore contain the complete 2D evaluation and data/training evidence for 3D, with the 3D evaluation rerun left to the regeneration command. This delivery cut is recorded in the decision log (D9) and the compliance matrix (S7, S8).

6 6 Discussion and limitations

What the study shows. The four constraint-injection points are not interchangeable. (i) The *data level* is the only point that gives exact satisfaction for free — vorticity/streamfunction in 2D and the Beltrami construction in 3D make $\nabla\cdot u \equiv 0$ and fixed energy balance properties of the *dataset*, verifiable to round-off (1.5×10^{-7}). (ii) *Model design* (curl of a vector potential) moves the constraint into the architecture and changes the geometry of the learning problem (smoother A-space targets), which is a fidelity/interpretability trade-off, not a free win. (iii) A *training-loss penalty* at a fixed weight $\lambda = 0.1$ failed to enforce divergence-free sampling within the budget — soft constraints need tuning. (iv) *Sampling procedures* operate on a fixed conditional model: they corrected energy bias but could not fix an underfit conditional distribution. The comparative conclusion (subject item S8) is therefore about *where constraints can be guaranteed vs where they merely nudge*: exactness lives in the data formulation and (up to solver round-off) in the Fourier projection; everything else trades fidelity for partial consistency.

Limitations. (1) The 3D benchmark is a synthetic solenoidal helical-field family, not MHD DNS: it reproduces divergence-freeness, helicity statistics and parameter-dependent spectra, but not magnetic-Reynolds-number dynamics, polarity reversals or non-Gaussian intermittency (subject items S3/S7 are partially met for this reason). (2) The 2D model is deliberately small (1.0 M parameters, 1500 steps); the negative 2D result is a statement about this budget, not about flow matching for turbulence. (3) One seed per configuration; no error bars on training variability. (4) The Re range (12.5–62.5 as defined by $\mathrm{Re} = A/(\nu\, k_f)$) is moderate; no inertial-range claims are made. (5) The 3D evaluation stage was interrupted by the delivery deadline (§5.3); its reported numbers will exist only after the regeneration run. (6) No societal impact beyond the general caveat that generative surrogate fields must not be presented as DNS data.

7 References

- Wang, S., Dou, Z., Shan, S., Liu, T.R. and Lu, L., 2026. Fundiff: Diffusion models over function spaces for physics-informed generative modeling. *Nature Communications*.
- Fukami, K. and Taira, K., 2023. Grasping extreme aerodynamics on a low-dimensional manifold. *Nature Communications*, 14(1), 6480.
- Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, R., Goldblum, M., Geiping, J. and Goldstein, T., 2024. Universal guidance for diffusion models. *ICLR 2024*.
- Bastek, J.H., Sun, W. and Kochmann, D., 2025. Physics-informed diffusion models. *ICLR 2025*.
- Solera-Rico, A. et al., 2024. β-Variational autoencoders and transformers for reduced-order modelling of fluid flows. *Nature Communications*, 15, 1361.
- Lipman, Y. et al., 2023. Flow matching for generative modeling. *ICLR 2023*. (rectified-flow formulation: Liu et al., 2023, *Flow straight and fast*, ICLR 2023)

8 Author contributions (AI models)

- **Harness agent** (ZCode CLI on `glm-5.3-flash`, provider zhipu): mission planning, experiment design, all experiment code (the delivered code), running/training/evaluation on the rented GPU, results analysis, report integration and §5–6.
- **Sub-agent** (`deepseek-v4-pro`, DeepSeek): drafted §1–4 prose from a detailed factual brief (0.0044 USD).
- **Sub-agent** (`glm-5.3-flash`, zhipu): drafted the compliance matrix, the NeurIPS checklist and the citation file scaffolding (0.0018 USD).
- Compute: 1× Runpod RTX 4090 (billed hourly); all code, figures and numbers machine-generated.