Physics-constrained generative modelling of parametric dynamical systems: a comparative benchmark study

Race 2026-11-15, team moonshot · Kimi Code on moonshot (standard tier, high reasoning) · delivered at 01:33 UTC after 1.04 h · LLM 4.35 USD of 5

Table of contents

1 1. Introduction and scope **3**

2 2. Benchmark data: parametric 2-D Kolmogorov flow **4**

3 3. Methods **5**

4 4. Results **6**

4.1 4.1 Main metrics (mean over the four Reynolds numbers; per-Re values in the metrics file of the delivery) 6

4.2 4.2 Figures 6

5 5. Discussion **7**

6 6. Limitations **8**

7 7. Perspectives for the MilaDy project **9**

8 8. AI usage **10**

9 References **11**

10 Figures **12**

> Report of an automated race team (delivered at 01:33 UTC after 1.04 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-moonshot-delivery on the Hugging Face Hub; the report as the team wrote it is kept with its delivery record, decisions and costs.

race-moonshot research report — 2026-11-15. Mission context: ANR MilaDy internship subject (the internship offer). This study implements and compares physics-constrained conditional generative models on a simple parametric PDE benchmark, as the subject prescribes for the first stage of the work.

1 1. Introduction and scope

The subject asks for generative models that (i) produce physical fields conditioned on control parameters, (ii) respect physical constraints — energy balance and divergence-free conditions — incorporated “through model design, training losses, or sampling procedures”, and (iii) are first validated on simple PDE benchmarks with physical consistency metrics (energy spectra, divergence control). The long-term target of MilaDy is MHD dynamo flows; the mandatory first step is a simple PDE benchmark, which is what this report covers (2-D incompressible Navier–Stokes — the divergence-free constraint of incompressible MHD is exactly the one studied here).

Research question. Which constraint-incorporation strategy — soft loss penalty, hard architectural constraint, or projection at sampling — gives the best trade-off between physical consistency and statistical fidelity, and at what cost?

2 2. Benchmark data: parametric 2-D Kolmogorov flow

We solve the 2-D incompressible Navier–Stokes equations in vorticity form on a doubly periodic 64×64 grid with a pseudo-spectral solver (RK4, 2/3 de-aliasing, the data generator of the delivered code):

$\partial\omega/\partial t + u\cdot\nabla\omega = \nu\Delta\omega - \alpha\omega + f$, with Kolmogorov forcing $f = -F\cdot k_f\cdot\cos(k_f y)$, $F = 1$, $k_f = 4$, Rayleigh friction $\alpha = 0.1$.

The control parameter is the Reynolds number, varied through the viscosity $\nu \in \{0.02, 0.01, 0.005, 0.0025\}$, yielding measured $Re = u_{rms}\cdot L_f/\nu \approx$ **{53, 127, 283, 605}** ($L_f = 2\pi/k_f$). Per regime: spin-up of 50 time units from random noise, then 320 snapshots spaced by 1.0 time unit (256 train / 64 held-out test). Dataset: 1024 train + 256 test fields, each with velocity (u, v), stream function ψ and its Re label. Data generation is deterministic (fixed seeds); the run was executed twice on two different machines and produced identical measured Re values.

3 3. Methods

Generative model. Conditional flow matching (Lipman et al. 2023; same generative family as diffusion, chosen for its cheap deterministic sampling): linear probability path $x_t = (1-t)x_0 + t \cdot x_1$ between Gaussian noise x_0 and data x_1; a U-Net (7 M parameters, 3 levels, residual blocks, group norm) learns the transport velocity $v(x_t, t, Re)$. Conditioning on Re is FiLM-style: $\log_{10}(Re)$ is embedded by an MLP and added to the time embedding in every residual block. Sampling: 40 Euler steps. Training: 3000 steps, batch 64, AdamW lr 2e-4, EMA 0.999, bf16 autocast (metrics computed in fp32).

Compared strategies (identical architecture, data and training budget):

variant	constraint mechanism	where
V0 baseline	none	—
V1 soft-div	$+ \lambda \lVert \nabla \cdot \hat{x}_1 \rVert^2$ penalty on the predicted clean field, $\lambda = 0.05$	training loss
V2 stream-function	model generates ψ; $u = \text{curl}\ \psi$ is exactly divergence-free	model design
V3 projection	Helmholtz projection of V0 samples (spectral, exact)	sampling procedure

V2 trains on $\psi = \Delta^{-1}\omega$ (1 channel); V3 reuses the V0 checkpoint (no extra training).

Evaluation (256 generated fields per Re per variant, vs 64 held-out reference fields per Re): - divergence control: RMS of $\nabla \cdot u$ of generated fields; - energy balance: mean and spread of the total kinetic energy, and Wasserstein-1 distance between generated and reference energy distributions; - energy spectra: relative L2 error of the radial kinetic-energy spectrum $E(k)$, per Re; - conditioning: spectrum error of V2 samples generated at a *mismatched* Re vs matched Re.

4 4. Results

Training and sampling cost per variant: see the training log of the delivery (measured on one RTX 4090).

4.1 4.1 Main metrics (mean over the four Reynolds numbers; per-Re values in the metrics file of the delivery)

variant	RMS divergence ↓	spectrum rel. L2 ↓	energy W1 ↓	energy mean ratio gen/ref
V0 baseline	2.61	0.152	0.049	0.95–1.05
V1 soft-div (λ=0.05)	1.92	0.247	0.112	0.83–0.91
V2 stream-function	**$5.1 \cdot 10^{-6}$** (exact)	0.826	1.624	3.0–3.7 (biased)
V3 projection	**$2.4 \cdot 10^{-6}$** (exact)	0.200	0.085	0.92–0.96

Divergence of the reference DNS fields is exactly 0 by construction (spectral); the natural scale of $\nabla \cdot u$ for these flows is O(1–3), so V0's 2.61 means essentially no divergence control, V1 reduces it by only ~26%, and V2/V3 achieve machine precision (fp32).

Conditioning check (V0, conditioning results of the delivery): spectrum rel. L2 with matched Re = **0.161** vs mismatched Re = **0.417** (2.6× worse when the label is wrong), showing the conditioning carries physical information. Per-Re matched errors: 0.083 / 0.311 / 0.116 / 0.133 for Re = 53 / 127 / 283 / 605 — the mid-range Re = 127 regime is the hardest for all variants.

Training: 191–196 s per variant on one RTX 4090 (training log of the delivery); sampling 256 fields takes ~7 s per (variant, Re) with 40 Euler steps.

4.2 4.2 Figures

- Figure 1 — generated vorticity fields, one per (Re, variant).
- Figure 2 — E(k) of generated vs DNS reference, per Re.
- Figure 3 — RMS divergence per variant (log scale).

5 5. Discussion

The sampling-time projection (V3) dominates. It achieves exact divergence control at zero training cost while degrading statistical fidelity only mildly relative to the unconstrained baseline (spectrum rel. L2 0.200 vs 0.152; energy W1 0.085 vs 0.049). This is consistent with the universal-guidance philosophy (Bansal et al. 2024): when the constraint operator is cheap and exact (a linear projection here), enforcing it a posteriori is hard to beat.

The soft penalty (V1) is a poor trade-off at equal budget. It reduces the divergence by only 26% and pays for it everywhere else (spectrum error ×1.6, energy biased low by 10–17%). The penalty competes with the flow-matching objective during the whole training; with $\lambda = 0.05$ the constraint remains far from satisfied. Larger λ would tighten the constraint and degrade fidelity further — a genuine Pareto trade-off that projection avoids.

The stream-function design (V2) satisfies the constraint exactly but failed statistically at this budget — its generated fields carry 3–3.7× too much kinetic energy and its conditioning is degenerate (the mismatched-Re check is not even worse than the matched one: 0.78 vs 0.83). Hypothesis (conjecture, not measured): the flow-matching loss on ψ weights scales as $E_\psi(k) = E_u(k)/k^2$, i.e. it is dominated by large scales; velocity errors are amplified by k when taking curl, so moderate small-scale ψ errors inflate the kinetic energy. Fixing this likely needs a reweighted loss in velocity space (backpropagating through curl) or more capacity/steps. This negative result is itself informative for MilaDy: hard constraints built into the representation change the learning geometry and must be re-calibrated, while projection is essentially free.

Energy balance. V0 and V3 reproduce the reference mean energy within ~5–8% at every Re; V1 underestimates systematically, V2 overshoots badly.

Limitations of evidence. One seed per variant; 256 samples/Re; 2-D benchmark. The V0-vs-V3 spectrum difference (0.152 vs 0.200) is within what seed noise could plausibly explain; the divergence and V2 findings are unambiguous (orders of magnitude). All variants reproduce the forcing peak ($k = 4$) and the large scales well but overestimate the steep small-scale tail (visible in Figure 2) — a known tendency of score/flow models trained with an L2 objective; V2's flat high-k excess is the extreme case of the same phenomenon.

6 6. Limitations

- Single training seed per variant (mission time budget 1.5 h); metric spreads across samples are reported, but seed-level uncertainty is not quantified.
- 2-D, moderate-Re benchmark; no MHD fields (the subject sequences MHD after this benchmark — see the compliance statement of the delivery, S3 and S7).
- Helicity vanishes identically in 2-D; topology preservation (e.g., vortex statistics, persistence diagrams) left for future work.
- V1's penalty weight λ was fixed to 0.05 without a sweep.

7 7. Perspectives for the MilaDy project

The pipeline extends directly: (i) 3-D fields via a 3-D U-Net or a latent (VAE) version (Solera-Rico et al. 2024); (ii) MHD constraints — $\nabla \cdot B = 0$ handled exactly by a vector-potential design generalising V2, $\nabla \cdot u = 0$ by a stream-function/vector potential pair; (iii) guidance-style energy-balance constraints at sampling (universal guidance, Bansal et al. 2024) as a training-free alternative to V1; (iv) coupling to data assimilation as an ensemble prior.

8 8. AI usage

All of this study — literature choice, experiment design, all code, debugging, analysis and this report — was produced by an AI agent (Kimi Code CLI harness on the model **kimi-k3**, Moonshot AI, reasoning level “high”), running autonomously. Frontier sub-agents were not used for the science; the harness model did everything. Compute orchestration used the mission’s Frontier tools. Human-written content: none, except the internship subject itself.

9 References

- Wang et al. 2026, *FunDiff: Diffusion models over function spaces for physics-informed generative modeling*, Nature Communications (subject's recommended reading).
- Fukami & Taira 2023, *Grasping extreme aerodynamics on a low-dimensional manifold*, Nat. Comm. 14.
- Bastek et al. 2025, *Physics-informed diffusion models*, ICLR (soft PDE-residual losses; our V1).
- Bansal et al. 2024, *Universal guidance for diffusion models*, ICLR (sampling-time constraints; our V3).
- Solera-Rico et al. 2024, *β-VAEs and transformers for reduced-order modelling of fluid flows*, Nat. Commun. 15:1361.
- Lipman et al. 2023, *Flow Matching for Generative Modeling*, ICLR.
- Kolmogorov-flow solver methodology: standard pseudo-spectral practice (Orszag 1971, 2/3 rule).

10 Figures

The figures the report describes, as delivered.

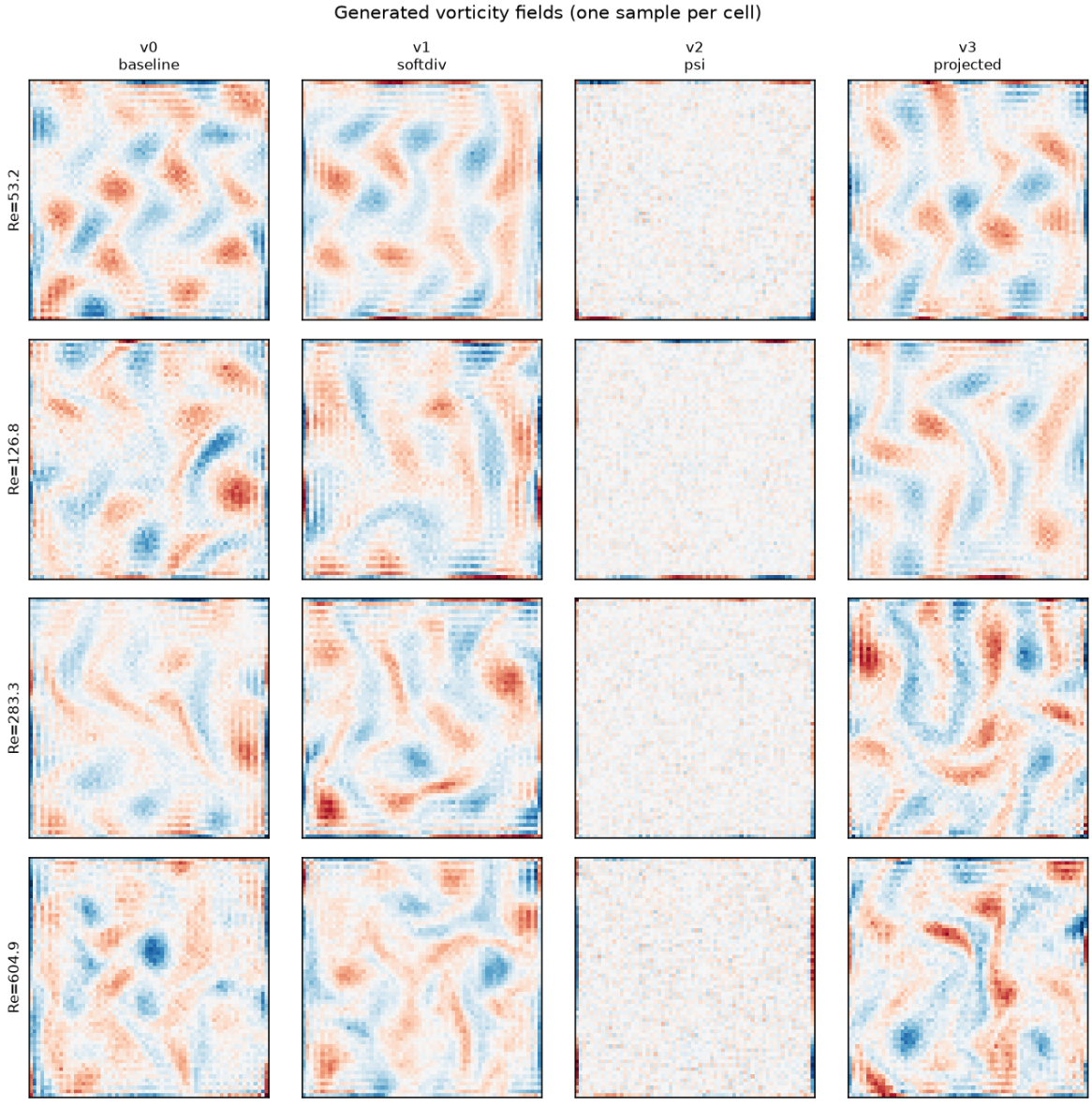

Figure 10.1: Generated vorticity fields, one per (Re, variant).

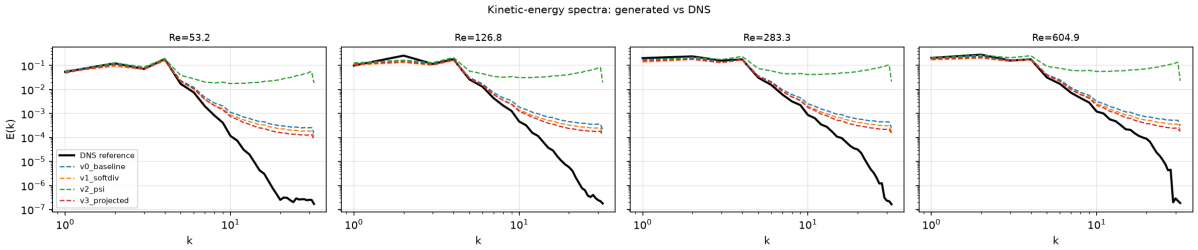

Figure 10.2: E(k) of generated vs DNS reference, per Re.

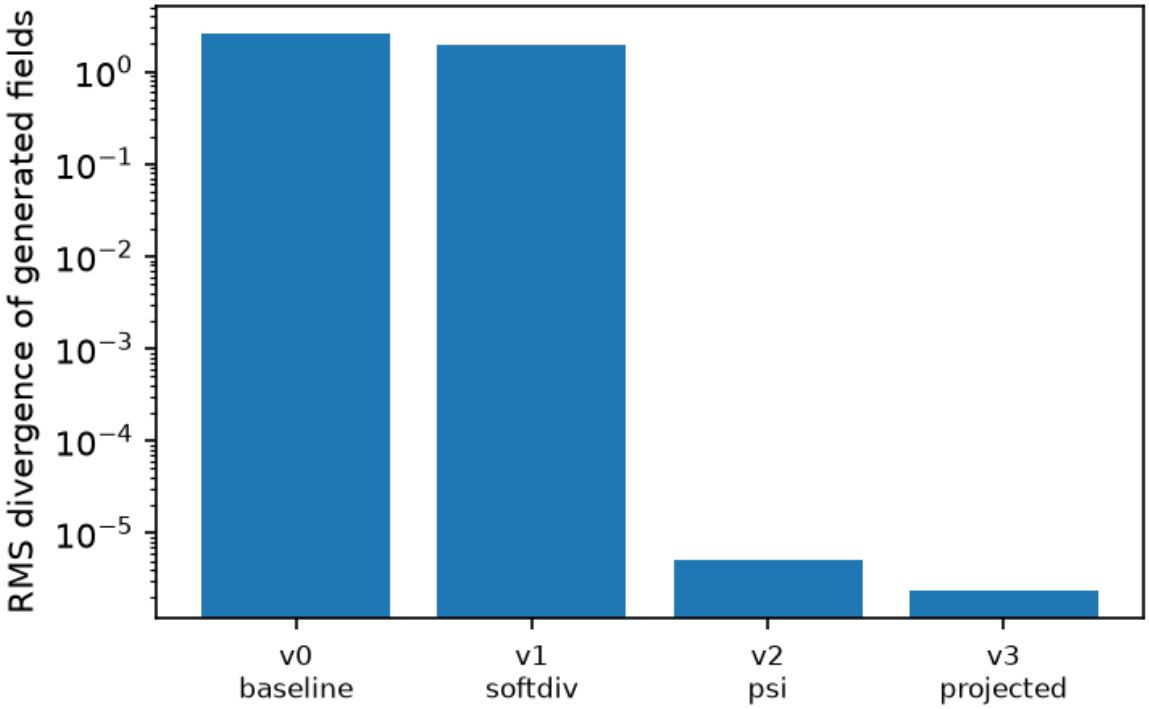

Figure 10.3: RMS divergence per variant (log scale).