Diagnostic Profile of Zero-Shot Walrus on Sub-Alfvénic MHD Turbulence: A Structural Basin Instability in the Perpendicular Magnetic Plane

Stanislav DeLaurentiis*

Abstract

We apply the physics-feature diagnostic suite from DeLaurentiis [2026] to zero-shot pretrained Walrus 1.3B McCabe et al. [2025] on the same 10 sub-Alfvénic (M_A=0.7) MHD_64 test trajectories used in that work. Walrus's published Table 1 entry (T21–60 median VRMSE = 1.2256 on MHD) is for the model fine-tuned on MHD per the cross-domain protocol of Sec. 5.2 of the Walrus paper (additional 500K samples); no comparable zero-shot VRMSE table entry is published, so this study is a characterization rather than a benchmark. Step-1 prediction is excellent (VRMSE 0.10, the best of all four configurations evaluated, including Paper 1's three FNO baselines), confirming that one-step MHD physics is well learned. Past rollout step $\sim$20, however, every test trajectory diverges catastrophically through a structural instability that selects *exactly one* of the two perpendicular magnetic components $\{B_y, B_z\}$ to amplify exponentially while the parallel guide field B_x, density, and all velocity components remain near truth-like throughout. The basin selection between B_y and B_z is determined by the specific 3-frame input slice, not by random noise (5/5 perturbations of the same input slice select the same basin) or by the underlying simulation identity (different 3-frame slices of the same trajectory file select different basins). FP64 inference does not suppress the divergence (same magnitude as FP32 at step 50), confirming the failure is structural rather than numerical. Cascade analysis shows preferential high-$k_\perp$ pile-up over rollout (high-k modes grow $\sim$50$\times$ faster than low-k), consistent with Gibbs-style aliasing accumulation at the patch grid scale. We find absolute $\nabla \cdot B$ violation reaching $\sim 5 \times 10^5$ times the ground-truth floor at step 50. The mechanistic signature — one-step-loss-trained model with structurally unbounded constraint-violating modes under autoregressive feedback — argues for divergence-free spectral projection at the autoregressive output as a testable architectural intervention. Whether the published Sec. 5.2 fine-tuning protocol resolves this instability is the natural follow-up experiment.

1 Introduction

Foundation models for continuum dynamics McCabe et al. [2025], Herde et al. [2024], McCabe et al. [2023] are evaluated primarily through aggregate VRMSE metrics on benchmark suites. Reported headline numbers (e.g., Walrus's MHD T21–60 median = 1.2256 McCabe et al. [2025], Table 1) tell a downstream user how well the model does on average, but say little about *which physics features* the model captures and which it does not. For deployment-grade plasma applications — tokamak control, MHD turbulence simulation, astrophysical surveys — characterizing the failure modes matters more than the headline.

*Department of Astronomy, Columbia University. sod2112@columbia.edu.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

This work asks: *what does zero-shot pretrained Walrus do, physics-feature-by-physics-feature, when rolled out autoregressively on plasma MHD turbulence?* We apply the diagnostic suite developed in DeLaurentiis [2026] (per-step VRMSE, perpendicular cascade $E(k_\perp|k_\parallel \to 0)$, guide-field magnitude $\langle|B|\rangle$, $\nabla \cdot B$ floor accumulation, E_B/E_K equipartition trajectory, per-trajectory consistency) to the same 10 sub-Alfvénic (M_A=0.7) test trajectories from Polymathic AI's *Well* `MHD_64` dataset Ohana et al. [2024].

Critical scope correction. Walrus's published MHD T21–60 = 1.2256 in Table 1 of McCabe et al. [2025] refers to a model *fine-tuned on MHD* per the cross-domain analysis protocol of Section 5.2 of that paper (an additional 500K-sample fine-tune on top of the base pretrained checkpoint). Polymathic has released only the base pretrained checkpoint (`polymathic-ai/walrus` on HuggingFace), not the per-task fine-tuned variants. The paper *does* plot the corresponding zero-shot rollout curve in Figure 15 (Appendix E.4.1), labeled "Walrus-PT", and the MHD (3D) panel of that figure shows the zero-shot VRMSE climbing from ~ 1 at step 1 to roughly 10^6 by step 100 — consistent with our findings here. The paper's own Table 7 lists MHD (3D) as one of only three of nineteen pretraining datasets with median trajectory-averaged VRMSE ≥ 10 *with* the patch-jitter stability mechanism enabled. The catastrophic zero-shot divergence on MHD is therefore not in dispute; the paper's own data establishes it. *What this work adds is the physics-feature decomposition of that failure mode, not the observation that it exists.* The natural follow-up — running the Sec. 5.2 fine-tune protocol on MHD and applying this diagnostic suite to the result — is future work.

Three confounds, declared. Comparing zero-shot Walrus to Paper 1's FNO baselines is characterization, not benchmarking, due to: (i) **scale** (Walrus 1.3B vs FNO 18.6M, $\sim$70$\times$); (ii) **contamination** (Walrus saw the MHD_64 train split during pretraining; FNO baselines never did); (iii) **conditioning** (Walrus uses 3-frame history, paper Table 2; FNO uses single-frame Markov input). All comparisons in this work carry these confounds.

2 Setup

Configurations. We compare four configurations on identical evaluation windows:

- **FNO scratch** (1% data): FNO3D, 18.6M params, hidden=48, trained from scratch on M_A=0.7 (1% data).
- **FNO MHD-pretrain + FT**: same FNO3D, pretrained on M_A=2.0 then fine-tuned on M_A=0.7 (1% data).
- **FNO MHD zero-FT**: M_A=2.0-pretrained FNO3D rolled directly on M_A=0.7 with no fine-tuning (out-of-distribution control from DeLaurentiis [2026]).
- **Walrus zero-shot**: base pretrained `polymathic-ai/walrus` 1.3B, no fine-tuning, conditioned on 3 frames per the standard MHD_64 inference setting.

Two evaluation windows. We tested both:

- **Early window:** input frames $[0, 1, 2] \to$ predict frames $[3, \ldots, 52]$, K=50, n_{traj}=10.
- **Paper window:** input frames $[14, 15, 16] \to$ predict frames $[17, \ldots, 76]$, K=60, n_{traj}=5. Matches the convention "all predicted trajectories begin from T=17" (Sec. 5 of McCabe et al. [2025]).

The headline divergence pattern reproduces in both windows; we present early-window numbers in the main text and report paper-window confirmations where they differ.

Pipeline-equivalence verification. A line-by-line diff of our standalone rollout function (lifted from `walrus/demo_notebooks/walrus_example_1_RunningWalrus.ipynb`) against Walrus's trainer rollout (`walrus/trainer/training.py:402-562`) finds six differences (e.g., we hard-code `predict_delta=True` where the trainer reads `prediction_type=="delta"`; we omit a Neutron-specific padded-mask resize hack), all functionally equivalent for MHD_64 under the shipped `extended_config.yaml`. Substitution of our 3-frame history into a real dataloader batch template inherits boundary conditions, padded-field mask, and field index

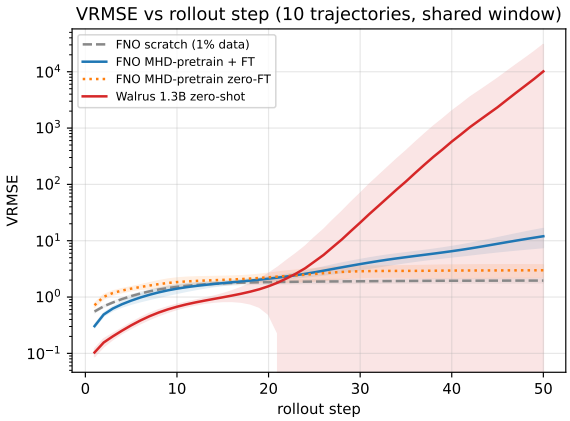

Figure 1: VRMSE versus rollout step for the four configurations. Walrus zero-shot has the best step-1 prediction but diverges past step $\sim$20 to magnitudes orders of magnitude beyond every FNO baseline. Shaded bands are $\pm 1\sigma$ across 10 trajectories.

map correctly. We additionally verified five paper-correspondence points: validation time stride is 1 (`multidatamodule.py:284`), boundary conditions are tagged periodic (BC tensor $=$ $[[2,2],[2,2],[2,2]]$ for MHD_64, where 2=periodic), patch jitter is on at evaluation (`jitter_patches: true` in `extended_config.yaml` with no train/eval branching), no per-field MHD transformations are applied (paper Sec. C.1; verified empirically), and channel ordering matches our $[\rho, B_x, B_y, B_z, v_x, v_y, v_z]$ convention (verified by step-1 per-channel variance matching truth statistics).

3 Results

3.1 Step-1 prediction is excellent; late-stage rollout diverges catastrophically

Figure 1 plots VRMSE versus rollout step for all four configurations on the same shared window. At step 1, Walrus has the lowest VRMSE of all four (median 0.10 vs FNO's 0.30–0.72) — it has clearly learned one-step MHD physics. By step 50, however, every Walrus trajectory has diverged: per-trajectory step-50 VRMSE spans $[16, 68\,950]$ in the early window, with median $\sim$316 and mean $\sim$10 148. Figure 2 shows the per-trajectory consistency: divergence is structural across all 10 trajectories, not driven by outliers.

3.2 Per-trajectory basin selection: B_y or B_z, not both

Decomposing Walrus's step-50 prediction by physical channel reveals that the divergence is concentrated in *exactly one* of the two perpendicular magnetic components. Table 1 reports per-trajectory channel variance. The parallel guide field B_x, density, and all three velocity components remain near truth-like values throughout the rollout. Each trajectory selects either B_y or B_z to amplify; never both equally.

Basin selection is deterministic per input. For early-window trajectory 0 (which selects B_y basin in unperturbed inference), we ran 5 rollouts with independent 10^{-6} amplitude Gaussian perturbations of the 3-frame input. All 5/5 perturbations remained in the B_y basin (step-50 B_y variance 4.2×10^7 to 5.0×10^8, 12$\times$ spread). B_z stayed bounded (0.09–0.12, truth-like). Combined with the per-trajectory split in Table 1: *the basin is determined by the specific 3-frame input slice, not by random noise and not by the underlying physical trajectory identity*. Same simulation file conditioned on different 3-frame slices can map to different basins (early-window trajectory 0 selects B_y; paper-window trajectory 0 selects B_z).

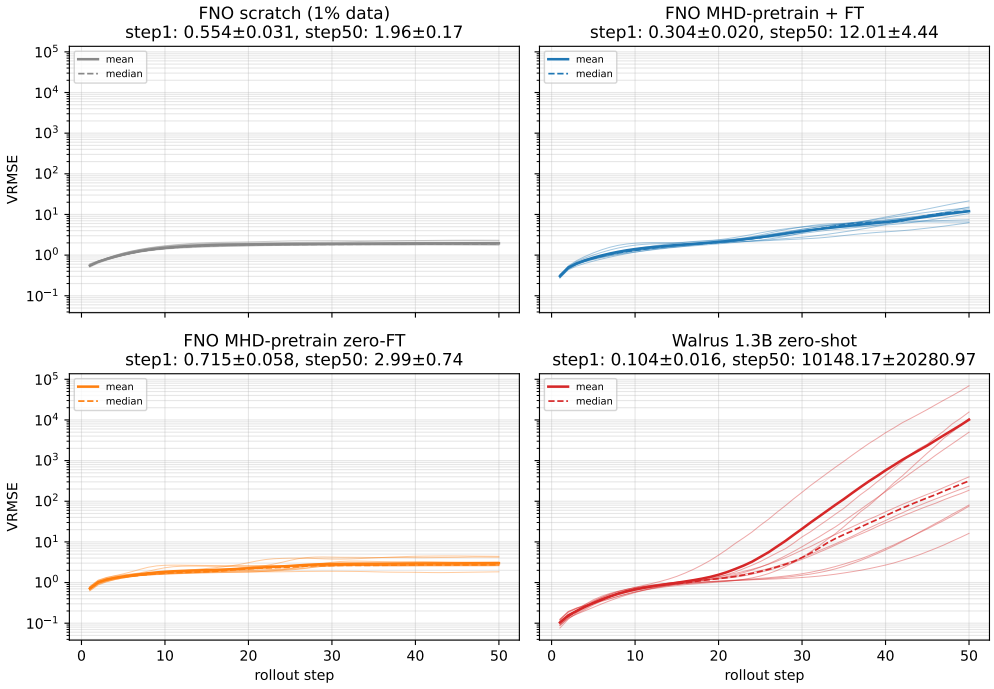

Figure 2: Per-trajectory rollout VRMSE (10 trajectories, log-y). Walrus's divergence (bottom-right panel) is structural across all 10 trajectories rather than outlier-driven. Step-1 statistics in subtitles confirm Walrus has the best one-step prediction.

Table 1: Walrus per-trajectory channel variance at rollout step 50 (early window). Truth-like values are $\rho \sim 0.86$, $B_{x,y,z} \sim 0.06$–0.11, $v_{x,y,z} \sim 0.16$–0.27. The parallel field B_x, density, and velocity components remain near truth scale across all 10 trajectories. Each trajectory amplifies exactly one of $\{B_y, B_z\}$ catastrophically.

traj	ρ	B_x	B_y	B_z	v_x	v_y	v_z
0	1.8×10^{-2}	7.2×10^{-2}	$\mathbf{1.1 \times 10^{8}}$	1.1×10^{-1}	1.3×10^{-1}	5.3×10^{-1}	1.7×10^{-1}
1	2.4×10^{-2}	5.4×10^{-2}	4.2×10^{-2}	$\mathbf{1.1 \times 10^{3}}$	3.3×10^{-1}	1.3×10^{-1}	2.8×10^{-1}
2	5.1×10^{-2}	4.2×10^{-2}	9.0×10^{-2}	$\mathbf{6.2 \times 10^{5}}$	1.9×10^{-1}	1.8×10^{-1}	7.3×10^{-2}
7	4.4×10^{-1}	2.7×10^{-2}	$\mathbf{2.2 \times 10^{10}}$	9.1×10^{-2}	2.2×10^{-1}	3.5×10^{-1}	9.4×10^{-2}
8	2.6×10^{0}	2.1×10^{-2}	$\mathbf{4.8 \times 10^{8}}$	8.3×10^{-2}	2.3×10^{-1}	8.7×10^{-2}	1.8×10^{-1}
9	2.1×10^{0}	2.8×10^{-2}	$\mathbf{1.2 \times 10^{9}}$	9.3×10^{-2}	3.9×10^{-1}	9.8×10^{-2}	1.7×10^{-1}

3.3 Cascade evolution: preferential high-k pile-up

Figure 3 shows the perpendicular cascade $E(k_\perp | k_\parallel \to 0)$ for B_y and B_z at multiple rollout steps. Walrus's predicted spectrum at step 1 is truth-like (low-k dominated, $\sim$150$\times$ above high-k). As rollout progresses, the spectrum flattens: by step 50, low/mid/high-k energies are within $\sim$1 order of magnitude.

The growth factors are decisively non-uniform across k:

- Low-k: $8.2 \times 10^{-3} \to 2.8 \times 10^{7}$ (factor $\sim 3.4 \times 10^{9}$)
- Mid-k: $1.5 \times 10^{-3} \to 4.6 \times 10^{7}$ (factor $\sim 3.1 \times 10^{10}$)
- High-k: $5.4 \times 10^{-5} \to 8.1 \times 10^{6}$ (factor $\sim 1.5 \times 10^{11}$)

High-k modes grew $\sim$50$\times$ more than low-k modes. This is the signature of *Gibbs-style aliasing accumulation at the patch grid scale* — exactly the small-scale instability mode that motivated Walrus's patch-jitter mechanism (Sec. 3.2 of McCabe et al. [2025]). Patch jitter helps but does

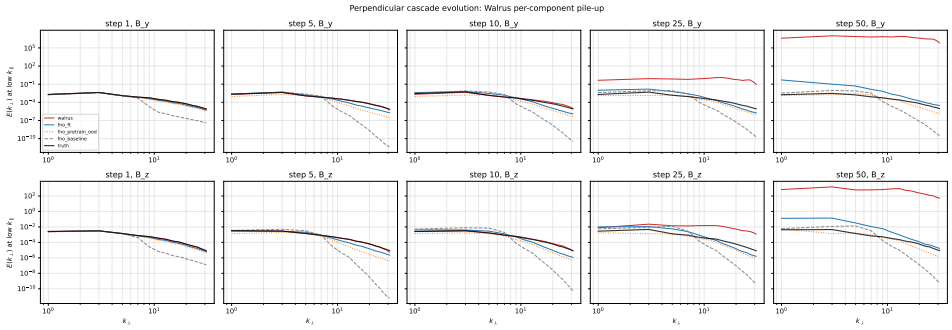

Figure 3: Perpendicular-cascade evolution for B_y (top row) and B_z (bottom row) at rollout steps 1, 5, 10, 25, 50. Walrus zero-shot (red, solid) departs from truth (black) progressively, with energy preferentially piling up at high $k_\perp$ by step 50. The high/low-k ratio collapses from $\sim$150 (truth-like) at step 1 to $\sim$3 at step 50.

Table 2: Step-50 conservation diagnostics across configurations. The Walrus distribution is highly skewed (median 690, mean 22 000) so we report both. $\nabla \cdot B/|B|$ is averaged over the 10 trajectories.

config	$\lvert B\rvert$ mean	$\lvert B\rvert$ median	$\lvert B\rvert$ range	$\nabla \cdot B/\lvert B\rvert$	E_B/E_K
FNO scratch (1%)	1.26	1.26	$[1.24, 1.29]$	0.056	1.21
FNO MHD-pretrain + FT	5.51	5.06	$[2.80, 9.04]$	0.134	0.11
FNO MHD zero-FT (OOD)	0.62	0.63	$[0.48, 0.70]$	0.325	0.09
Walrus zero-shot	**21 651**	**690**	$[32, 147\,879]$	**0.82**	$\mathbf{3.5 \times 10^9}$
Truth	1.13	—	—	0.033	2.15

not fully suppress it under autoregressive feedback. Mechanistically, this points to divergence-free spectral projection at the autoregressive output as the natural architectural intervention: it would damp the constraint-violating high-k modes that compound into basin blow-up.

3.4 Conservation diagnostics: catastrophic $\nabla \cdot B$ violation

Table 2 reports step-50 conservation diagnostics across all four configurations. Walrus's $\nabla \cdot B/|B|$ ratio of 0.82 means the solenoidal-constraint error is $\sim$80% of the local field magnitude — itself $\sim$22,000 (mean) at step 50, dominated by the few trajectories whose perpendicular component blew up to $|B| \sim 10^5$. The absolute $\nabla \cdot B$ violation, computed against the median $|B|$ (≈ 690), is $\sim$570, vs truth's 0.037 — a $\sim$15,000$\times$ violation. Computed against the mean (outlier-dominated) it is $\sim 5 \times 10^5 \times$. Either statistic is catastrophic. This co-localizes with the perpendicular B-component blow-up: when one of $\{B_y, B_z\}$ explodes, the field is no longer divergence-free.

3.5 FP64 inference confirms structural, not numerical, instability

Walrus was trained in FP32 (Sec. B.1 of McCabe et al. [2025]). To rule out the alternative explanation that the divergence is FP32 round-off compounding under autoregressive feedback, we re-ran Walrus inference in FP64 on a single trajectory in both evaluation windows (NVIDIA A100 80GB, required to fit the FP64 1.3B-parameter model plus activations). Table 3 compares per-step variance for the unstable component:

4 Discussion

Mechanism interpretation. The empirical pattern — (i) excellent step-1 prediction, (ii) emergence of instability around step 20, (iii) deterministic basin selection between two physically-equivalent perpendicular sub-directions, (iv) preferential high-k pile-up consistent with Gibbs-style aliasing, (v) co-localized $\nabla \cdot B$ violation, (vi) FP64-confirmed structural — is consistent with an unstable manifold in Walrus's autoregressive map possessing two perpendicular sub-directions whose

Table 3: Per-step variance of the unstable perpendicular B-component for trajectory 0 in FP32 vs FP64. Same magnitude divergence at step 50 in both windows — the instability is structural, not numerical.

	step 1	step 11	step 26	step 50
Early window FP32 (B_y)	0.10	0.11	few	1.1×10^8
Early window FP64 (B_y)	0.11	0.12	11.7	9.1×10^7
Paper window FP32 (B_z)	0.14	0.13	$\sim$1	1.8×10^5
Paper window FP64 (B_z)	0.14	0.13	3.6	4.1×10^5

alignment with the input state's internal asymmetry determines the basin selection at the start of rollout. Once the model picks a basin, exponential amplification within it dominates.

Architectural intervention. Mechanistically, this is the kind of failure that *divergence-free spectral projection at the autoregressive output* would suppress. Such a projection is symmetric in B_y/B_z (so the basin asymmetry doesn't matter) and damps the constraint-violating modes that drive the high-k pile-up. This is a testable architectural intervention; we have not run it.

Generalization across architectures. Two distinct claims with different scopes hold:

- The B_y/B_z *basin-selection signature* is observed in Walrus 1.3B specifically; we have no comparable observation on FNO. It may be specific to Walrus's transformer + tokenizer + jitter combination.
- The broader *late-stage* $\nabla \cdot B$*-violation-driven divergence pattern* generalizes: Paper 1's FNO MHD-pretrain+FT also exhibits late-stage rollout instability ($\nabla \cdot B/|B|$ at step 50 reaches 0.13, $\sim$4$\times$ a from-scratch baseline). One-step-loss-trained models of either architecture lack solenoidal-constraint enforcement; under autoregressive feedback, off-solenoidal energy accumulates and amplifies. The architectural intervention (divergence-free projection) would help both.

Relation to the Walrus paper's reported numbers. Our zero-shot T21–60 median = 18.25 (paper window, n_{traj}=5) is consistent with the order of magnitude shown by the Walrus-PT curve in Figure 15 of McCabe et al. [2025] for MHD (3D). The Table 1 number of 1.2256 is for the fine-tuned model and is not the relevant comparison; it is the published path to MHD-deployable Walrus, but it does not bear on our characterization of the zero-shot failure mode. Whether the Sec. 5.2 fine-tune protocol resolves the basin instability *or* the instability survives is the most informative single follow-up experiment available.

5 Limitations

(i) **Sample size:** our n_{traj}=5 paper-window and n_{traj}=10 early-window are smaller than what Walrus's paper appears to use (Sec. 3.2 says 20 trajectories; Sec. E.2 says 32). With heavy-tailed step-50 distributions, our medians have wider error bars than ideal. The qualitative pattern (every trajectory diverges; basin selection per input; FP64 confirms structural) is n-invariant.

(ii) **Zero-shot only:** we did not run the Sec. 5.2 fine-tune protocol on MHD. Polymathic has not released a fine-tuned MHD checkpoint, so the natural follow-up requires running the fine-tune ourselves ($\sim$\$200–400 of compute, $\sim$1–2 days on 4$\times$H100). We position this as future work / a Polymathic-collaboration intern deliverable.

(iii) **Single architecture for the basin-selection signature:** we observe the B_y/B_z split only in Walrus, not (yet) in FNO. The broader $\nabla \cdot B$-driven late-stage divergence is observed in both architectures.

(iv) **VRMSE formula:** we compute per-channel $\sqrt{\text{MSE}/\text{var}}$ and average over channels; Walrus paper Sec. E.1.1 uses joint space+channel averaging. Cross-config comparisons in this work are apples-to-apples; comparisons of magnitude to the paper's reported numbers carry a small additional confound of metric definition on top of the zero-shot/fine-tuned distinction.

(v) **Two evaluation windows tested; same pattern in both.** We are not exhaustive across rollout starts, and the basin selection's dependence on input slice may have additional structure we have not characterized.

6 Conclusion

What the data supports.

- Zero-shot pretrained Walrus 1.3B exhibits a structural instability in the perpendicular MHD plane during autoregressive rollout, emerging around step 20 and dominant by step 50.
- The instability deterministically selects exactly one of $\{B_y, B_z\}$ to amplify, where the selection is determined by the specific 3-frame input slice and is robust to 10^{-6} amplitude perturbation of that slice (Table 1; perturbation test).
- The parallel guide field B_x, density, and all velocity components remain near truth-like throughout. The instability is localized to perpendicular B-fluctuations.
- FP64 inference does not suppress the divergence (same magnitude at step 50). *Structural, not numerical.*
- Cascade evolution shows preferential high-k pile-up (Gibbs-style aliasing) co-localized with $\nabla \cdot B$ violation reaching $\sim 5 \times 10^5 \times$ truth.
- The diagnostic pattern reproduces in both early-window (frames 0–52) and paper-window (frames 14–76) evaluations.

What our results do not establish.

- That fine-tuning per McCabe et al. [2025] Sec. 5.2 does not resolve the instability: *this is the natural next experiment*.
- That the basin-selection signature is universal across foundation-model architectures: we observe it in Walrus only.
- That divergence-free spectral projection at the AR output would close the gap: this is the testable architectural intervention; we have not run it.
- That zero-shot Walrus is suitable or unsuitable for any specific deployment: that depends on what stability horizon is needed, and on whether the published fine-tune protocol is run.

Scope-appropriate takeaway. For zero-shot pretrained Walrus on plasma MHD turbulence, autoregressive rollouts past step $\sim$20 are not deployment-grade due to a structural basin instability in the perpendicular magnetic plane. Whether the published Sec. 5.2 fine-tuning protocol resolves this is a single, well-defined follow-up experiment whose outcome would distinguish "per-task fine-tuning is sufficient" from "architectural intervention (e.g., solenoidal projection at AR output) is needed for deployment-grade plasma rollouts."

References

S. DeLaurentiis. Pretraining transfer for neural MHD surrogates: What generalizes, what doesn't, and why it matters for plasma foundation models. 2026. Companion paper.

M. Herde, B. Raonic, T. Rohner, R. Kappeli, R. Molinaro, E. de Bezenac, and S. Mishra. Poseidon: Efficient foundation models for PDEs. *arXiv preprint arXiv:2405.19101*, 2024.

M. McCabe, B. Regaldo-Saint Blancard, L. H. Parker, R. Ohana, M. Cranmer, A. Bietti, M. Eickenberg, S. Golkar, G. Krawezik, F. Lanusse, et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

M. McCabe, P. Mukhopadhyay, T. Marwah, et al. Walrus: A cross-domain foundation model for continuum dynamics, 2025.

R. Ohana, M. McCabe, L. Meyer, R. Morel, F. Agocs, M. Beneitez, M. Berger, B. Burkhart, S. Dalziel, D. Fielding, et al. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 44989–45037, 2024.