POST-AGI PHYSICAL WORLD FOUNDATION MODELS: EVENT-CONDITIONAL DYNAMICS, TOPOLOGY TRANSFER, AND RISK-LIMITING GUARANTEES

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical decision-making is a limit case for post-AGI foundations: compounding errors, distribution shift, rare events, and institutional demand for auditability (Amodei et al., 2016). Time-series foundation models establish "pretrain then transfer" feasibility for forecasting (Das et al., 2023; Ansari et al., 2024), but high-stakes dynamical control adds two requirements that do not appear in benchmark-only regimes: (i) event-conditional, topology-aware world models that represent cross-stage propagation and delayed effects and (ii) evaluation contracts that connect uncertainty to deploy or abstain decisions. We propose Industrial Dynamics Foundation Models as a testable blueprint with three concrete components: unified tokenization of multi-rate telemetry and sparse operational events; graph-conditioned topology adapters that transfer a frozen backbone across plants; and shadow-mode guarantees built from regime-aware conformal calibration (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021; Stankeviciute et al., 2021). Guarantees become operational via a decision gate: recommend interventions only when safety constraints hold for all trajectories in a calibrated uncertainty set, and otherwise abstain. The full proposal is falsifiable with modest infrastructure: logged telemetry, event streams, and a process graph compatible with common industrial abstractions (OPC Foundation, n.d.; International Society of Automation, n.d.).

1 MOTIVATION

Industrial production lines instantiate post-AGI stressors that break purely score-driven development: partial observability, sparse regime changes, topology-coupled propagation, and delayed effects. In these settings, it is not enough to be accurate on average. A system must transfer across plants under bounded compute and data and must supply auditable reasons to act or abstain.

This paper frames industrial dynamics as a clean technical foundations arena: dynamics where the world pushes back, and where institutions demand decision-linked uncertainty. This connects directly to the workshop theme of resilient research directions: a post-AGI evaluation contract should remain meaningful as models scale.

2 RELATED WORK AND GAP

Time-series foundation models demonstrate broad transfer across forecasting tasks (Das et al., 2023; Ansari et al., 2024). Graph-based spatiotemporal models capture topology and propagation in structured domains (Li et al., 2017; Yu et al., 2018). Robust control and model predictive control provide a language for constraint satisfaction under uncertainty (Mayne et al., 2000; 2005).

The gap is a unified foundation-model direction that is simultaneously event-conditional, topology-aware, and paired with a deployment contract that yields bounded-risk decisions in shadow mode. Our contribution is that contract and the minimal modeling machinery needed to test it.

3 PROBLEM SETTING

Let a process line be a directed graph $G = (V, E)$. Each node $v \in V$ is a stage with latent state $x_t^{(v)} \in \mathbb{R}^{d_v}$ and observed telemetry $y_t^{(v)}$ arriving at multiple rates with missingness. Let $e_t \in \mathcal{E}$ be a sparse event stream and let $u_t \in \mathcal{U}$ be controllable inputs.

An Industrial Dynamics Foundation Model is a pretrained model that, given history and topology, predicts a distribution over H-step futures:

$$p_\theta(\tau_{t+1:t+H} \mid y_{\le t}, e_{\le t}, u_{\le t}, G),$$

where τ includes future telemetry and optionally event likelihoods.

Scope of "interventions". This tiny paper treats u_t in two ways: (i) as logged control inputs for world modeling and (ii) as candidate counterfactuals inside a simulator or a high-fidelity digital twin when evaluating the decision gate. When only logs exist, the paper still yields a fully testable contract for prediction, calibration, and abstention, without claiming closed-loop optimal control.

4 IDFM: MODEL CLASS, OBJECTIVE, AND TRANSFER

4.1 UNIFIED TOKENIZATION

We use a minimal tokenization that is implementable on real logs:

- Telemetry tokens: each measurement yields a tuple token $(v, s, \Delta t, z)$ with stage id v, sensor id s, time delta Δt, and value codeword z.
- Event tokens: each event yields $(v, \text{type}, \text{marker}, a)$ with stage id, event type, start or end marker, and optional attributes.
- Topology conditioning: stage ids and edge types parameterize propagation through adapters.

Events are treated as interventions to force regime-conditional learning rather than being filtered out as outliers.

4.2 TRAINING OBJECTIVE

Given a token stream $\mathbf{z}_{\le t}$ that interleaves telemetry and events, pretraining minimizes autoregressive negative log-likelihood:

$$\min_\theta \ \mathbb{E}\left[-\sum_{k=t+1}^{t+H} \log p_\theta(z_k \mid \mathbf{z}_{\le k-1}, G)\right] + \lambda\, \mathcal{L}_{\text{mask}}.$$

Event-conditional supervision is explicit: windows containing events are included and scored.

4.3 TOPOLOGY ADAPTERS FOR TRANSFER

Plants differ in graph structure, sensor sets, and stage ordering. We propose topology adapters as small trainable modules conditioned on G while freezing the pretrained backbone.

Frozen backbone. A sequence model f_θ maps tokens to hidden states. For each stage-local token, denote its backbone state by $h_t^{(v)}$.

Trainable adapters. For each stage v, learn a node adapter:

$$\tilde{h}_t^{(v)} = h_t^{(v)} + A_{\text{node}}^{(v)} h_t^{(v)}.$$

For each edge $(v \to w) \in E$ with edge type $\ell(v, w)$, learn an edge adapter:

$$m_{v\to w,t} = A_{\text{edge}}^{(\ell(v,w))}\, \phi(\tilde{h}_t^{(v)}), \quad \bar{m}_{w,t} = \text{AGG}\Big(\{m_{v\to w,t} : (v \to w) \in E\}\Big),$$

and inject $\bar{m}_{w,t}$ via a residual into $\tilde{h}_t^{(w)}$. Only node adapters, edge adapters, and a small output head g are tuned on the target plant.

Baselines. (B1) full fine-tuning of θ, (B2) head-only tuning, (B3) adapters without topology messages, (B4) a graph model trained from scratch on target plant logs (Li et al., 2017; Yu et al., 2018).

Falsifiable claim C1. Under a fixed target-plant calibration budget and wall-clock training cap, topology adapters outperform (B2) and (B3) on forecast NLL and event-conditional error and approach (B1) while using substantially fewer trainable parameters, and they outperform (B4) at small data where pretrained priors matter.

5 SHADOW-MODE GUARANTEES

The core critique of many "UQ in practice" proposals is that they do not define what is guaranteed, for which object, and under what conditions. We therefore specify the guarantee target and its limitations explicitly.

5.1 WHAT IS GUARANTEED

We aim for coverage of a trajectory set, but full joint coverage over all sensors, stages, and time steps is too strong and usually yields overly conservative sets. Instead, we target calibrated coverage for a risk-relevant score and turn that score into an auditable gate.

Define a nonconformity score $S(\tau, \hat{\tau})$ for an H-step forecast:

$$S = \max_{v \in V_{\text{crit}}} \max_{t < k \le t+H} \frac{|y_k^{(v)} - \hat{y}_k^{(v)}|}{\sigma^{(v)}}.$$

Split conformal produces q_α such that

$$\Pr\left(S \le q_\alpha\right) \ge 1 - \alpha$$

under the calibration assumptions (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). This is a guarantee on the score, not a claim that every coordinate is simultaneously covered. The decision gate below uses the score guarantee to bound worst-case violations on safety-relevant quantities.

5.2 CONSTRUCTING THE UNCERTAINTY SET

Given a model forecast $\hat{\tau}$ and a threshold q_α, define an induced set of plausible trajectories:

$$\mathcal{P}_\alpha = \Big\{\tau : S(\tau, \hat{\tau}) \le q_\alpha\Big\}.$$

Operationally, this is equivalent to an envelope around the critical sensors and stages. For deterministic forecasters, it yields interval tubes. For probabilistic forecasters, it is implemented by accepting simulated rollouts whose score is below q_α.

5.3 DECISION GATE

Let $c(\tau) \le 0$ be a safety constraint evaluated on trajectories. We define an auditable deployment rule:

$$\textsc{Deploy}(u) = 1 \;\text{ only if }\; \max_{\tau \in \mathcal{P}_\alpha(u)} c(\tau) \le 0.$$

If the gate abstains, the system reports the reason: which constraint or which regime calibration failed. This mirrors robust control logic where safe sets are certified before actuation, but here the set is data-calibrated in shadow mode (Mayne et al., 2000; 2005).

5.4 HANDLING TEMPORAL DEPENDENCE AND REGIME SHIFT

Classical conformal relies on exchangeability. Industrial time series violate this, so we adopt two practical remedies and treat failures as measurable signals rather than hidden assumptions.

Regime-aware calibration. We use stratified calibration by regime r defined by event type, product family, shift, or topology region:

$$q_\alpha^{(r)} = \text{Quantile}_{1-\alpha}\Big(\{S_i : i \in \mathcal{C}_r\}\Big).$$

The gate uses the regime-specific threshold for the current context. This is a testable claim about matched regimes, not a universal guarantee.

Block calibration for dependence. When using sliding windows, calibration examples overlap and are dependent. We therefore form calibration blocks (non-overlapping windows) and apply conformal to block-level scores, following the conformal time-series forecasting line of work (Stankeviciute et al., 2021). This makes the protocol explicit about how dependence is handled.

Falsifiable claim C2. In matched regimes, block and regime-aware conformal achieves near-nominal score coverage. Under regime mismatch or abrupt shifts, coverage violations rise and predict gate abstention before constraint violations occur.

6 A MINIMUM VIABLE BENCHMARK

To avoid proprietary dependence, we propose LineSim-H, a synthetic benchmark generator that isolates the claimed phenomena.

6.1 GENERATIVE PROCESS

Let $G = (V, E)$ be a DAG. Latent dynamics evolve with delayed coupling:

$$x_{t+1}^{(v)} = A_{r(t)}^{(v)} x_t^{(v)} + \sum_{(u\to v)\in E} B_{r(t)}^{(u\to v)} x_{t-d(u,v)}^{(u)} + C_{r(t)}^{(v)} u_t^{(v)} + \xi_t^{(v)}.$$

Events induce regime switches in $r(t)$ that change parameters and noise scale. Observations are partial and multi-rate:

$$y_t^{(v)} = M^{(v)} x_t^{(v)} + \epsilon_t^{(v)},$$

with sensor-specific sampling schedules, missingness bursts, and dropouts.

6.2 SHIFTS

(S1) changed event schedules and durations, (S2) topology edits including bypass edges and added buffer nodes, (S3) tightened action constraints.

Each shift is designed to stress one claim: event-conditional modeling, topology transfer, or gate robustness.

7 EVALUATION PROTOCOL

We pre-register the following plots:

- World modeling: forecast NLL and event-conditional RMSE.
- Transfer: performance vs calibration size for adapters, head-only, full tuning, and a scratch graph baseline.
- Guarantees: score coverage with confidence intervals; set width; gate violation rate and abstention rate.
- Budget curves: coverage and abstention vs calibration window size and adapter update frequency.

This makes the paper falsifiable even if closed-loop actuation is not evaluated.

8 FRONTIER POSITIONING

IDFM targets a post-AGI technical foundations question: can we scale general world modeling into domains where topology, interventions, and irreversibility dominate, and can we pair it with an evaluation contract that remains meaningful as models scale. The contribution is a standardized contract: event-conditional topology-aware world modeling with bounded-compute transfer, paired with shadow-mode, gate-based guarantees calibrated from data (Das et al., 2023; Ansari et al., 2024; Shafer & Vovk, 2008; Angelopoulos & Bates, 2021; Stankeviciute et al., 2021).

9 LIMITATIONS

Score coverage is not full joint coverage. When shifts are extreme, conformal sets widen and abstention rises. This is treated as a measurable outcome: we report time-to-recovery after recalibration and the abstention tradeoff. Topology adapters can fail when the target plant lies outside the pretrained support; this is detected by rising event-conditional error and coverage violations, which trigger abstention.

ETHICS STATEMENT

This proposal is evaluation-first and shadow-mode-first. Any deployment should include privacy-preserving telemetry handling, least-privilege access control, and explicit governance for when outputs may influence actuation.

REPRODUCIBILITY STATEMENT

All components are specified as implementable artifacts: token schema, objective, adapter definition, baselines, block and regime-aware conformal calibration, gate definition, benchmark generator, shift probes, and metrics.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv*, 2016. URL https://arxiv.org/abs/1606.06565. arXiv:1606.06565.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv*, 2021. URL https://arxiv.org/abs/2107.07511. arXiv:2107.07511.

Abdul Fatir Ansari, Lorenzo Stella, Brad Turkovic, Valentina Zantedeschi, Zhe Wang, et al. Chronos: Learning the language of time series. *arXiv*, 2024. URL https://arxiv.org/abs/2403.07815. arXiv:2403.07815.

Abhimanyu Das, Weihao Kong, Rajat Sen, et al. A decoder-only foundation model for time-series forecasting. *arXiv*, 2023. URL https://arxiv.org/abs/2310.10688. arXiv:2310.10688 (TimesFM).

International Society of Automation. Isa-95 standard overview. Web resource, n.d. URL https://www.isa.org/standards-and-publications/isa-standards/isa-95-standard. Accessed 2026-02-12.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv*, 2017. URL https://arxiv.org/abs/1707.01926. arXiv:1707.01926.

David Q. Mayne, James B. Rawlings, Christopher V. Rao, and Pierre O. M. Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000. doi:10.1016/S0005-1098(99)00214-9. URL https://www.sciencedirect.com/science/article/pii/S0005109899002149.

David Q. Mayne, Maria M. Seron, and Saša V. Raković. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*, 41(2):219–224, 2005. doi:10.1016/j.automatica.2004.08.019. URL https://www.sciencedirect.com/science/article/pii/S0005109804002353.

OPC Foundation. Opc unified architecture. Web resource, n.d. URL https://opcfoundation.org/about/opc-technologies/opc-ua/. Accessed 2026-02-12.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008. URL https://www.jmlr.org/papers/v9/shafer08a.html.

Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2d3f6b0f8d1b1b4d3b1f3c3c8a4e8e7a-Abstract.html.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv*, 2018. URL https://arxiv.org/abs/1709.04875. arXiv:1709.04875.